Received: 2 September 2019 | Revised: 28 October 2019 | Accepted: 23 December 2019

DOI: 10.1002/gepi.22278

Genetic Epidemiology | OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org | WILEY

RESEARCH ARTICLE

Genetic effect estimates in case-control studies when a continuous variable is omitted from the model

Ying Sheng[1] | **Chiung-Yu Huang**[1] | **Siarhei Lobach**[2] | **Lydia Zablotska**[1] | **Iryna Lobach**[1] | **for the Alzheimer's Disease Neuroimaging Initiative**

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

[2]Applied Mathematics and Computer Science Department, Belarusian State University, Minsk, Belarus

Correspondence
Iryna Lobach, Division of Biostatistics, Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94158.
Email: Iryna.lobach@ucsf.edu

Funding information
National Institutes of Health, Grant/Award Number: 5R21AG043710-02

Abstract

Large-scale genome-wide analyses scans on massive numbers of various cases and controls are archived in the genetic databases that are publically available, for example, the Database of Genotypes and Phenotypes (https://www.ncbi.nlm.nih.gov/gap/). These databases offer unprecscendented opportunity to study the genetic effects. Yet, the set of nongenetic variables in these databases is often brief. From the statistical literature, we know that omitting a continuous variable from a logistic regression model can result in biased estimates of odds ratios (OR), even when the omitted and the included variables are independent. We are interested in assessing what information is needed to recover the bias in the OR estimate of genotype due to omitting a continuous variable in settings when the actual values of the omitted variable are not available. We derive two estimating procedures that can recover the degree of bias based on a conditional density of the omitted variable given the disease status and the genotype or the known distribution of the omitted variable and frequency of the disease in the population. Importantly, our derivations show that omitting a continuous variable can result in either under- or over-estimation of the genetic effects. We performed extensive simulation studies to examine bias, variability, false-positive rate, and power in the model that omits a continuous variable. We show the application to two genome-wide studies of Alzheimer's disease.

KEYWORDS

Alzheimer's disease, bias, case-control study, odds ratio, omitted continuous variable

1 | INTRODUCTION

Recent advances in genotyping technology generated a variety of massive datasets that are archived in publicly available databases, including the Database of Genotypes and Phenotypes https://www.ncbi.nlm.nih.gov/gap/, the Cancer Genome Athlas https://portal.gdc.cancer.gov/, the UK Biobank https://www.ukbiobank.ac.uk/. These databases provide valuable information that can be used to improve our understanding of the genetic predisposition to complex diseases, such as cancer, diabetes, neurodegenerative disease. Such analyses of association can be designed to answer various questions, one of which is to identify the genetic variants and rank them according to the strength of the evidence for an association with the complex diseases. As a result, we can obtain valuable clues to the underlying aeteologic mechanisms of complex diseases. A commonly overseen complication is that omitting a variable from a logistic regression model can substantially bias the genetic effect estimates. We are interested in identifying conditions needed to recover bias in settings when the actual values of the omitted variable are not available to the researcher.

As discussed by many authors, including Gail, Wieand, and Piantadosi (1984), Neuhaus and Jewell (1996), Hauck, Neuhaus, Kalbfleisch, and Anderson (1991), Zeger, Liang, and Albert (1988), omitting variables associated with the disease can bias the odds ratio (OR) estimates. This is because the OR estimates reflect both the effect size and variability that is not explained by the model. Gail et al. (1984), Neuhaus and Jewell (1996), Zeger et al (1988) showed how the magnitude of bias in the estimate is a function of the true OR of the omitted variable in various settings. The correct adjusted OR estimate of the omitted variable, however, is rarely available because such an estimate would be based on a model that includes both the genetic and nongenetic variables. Hence we are interested in exploring what information available in the literature can recover bias in the estimates. We are also interested in assessing what information is needed to determine the directionality of the bias.

The setting considered in this paper is unique because we deal with three types of misspecifications at the same time. First, the data are collected using a retrospective design where the cases and controls are sampled independently from their respective populations, while the data are analyzed in a prospective logistic regression model. As shown in the seminal work by Prentice and Pyke (2011), we know that this aspect of misspecification does not result in based OR estimates because the OR can be estimated consistently from the retrospective likelihood-based methods. Secondly, the model is misspecified because a variable is omitted from the model. The second misspecification leads to the third, that is if the true model is logistic, the model that omits a variable might not be logistic.

The setting considered here is also unique in that the estimated effect size of a genetic variant is usually small to moderate, ranging between $-\log(1.5)$ and $\log(1.5)$ (Park et al., 2012). A few exceptions, however, have been reported in the literature. For example, in the context of Alzheimer's disease, the ApoE genotype is estimated to have an OR of 3.1 for heterozygous $\varepsilon 4$ genotype and 34.3 for homozygous $\varepsilon 4$ genotype (Kukull et al., 1979).

Our paper is organized as follows. We first perform a series of simulation studies to assess the bias due to omitting a variable empirically. The simulations are described in the Assessment of the Bias Due to Omitting a Variable (Section 2.1). Next, in Section 2.2 we derive the relationships between parameters of the reduced model where the variable is omitted and the parameters in the full model where the variable is included. We further conduct simulation studies described in Section 2.3 to assess how various pieces of information can contribute to the recovery of the bias. We show the application to the studies of Alzheimer's disease. And we conclude the paper by a brief discussion.

2 | MATERIALS AND METHODS

2.1 | Assessment of the bias due to omitting a variable

We first perform a series of simulation studies to assess potential bias, variance, mean squared error (MSE), false discovery rate (FDR), and power reduction when omitting an important continuous variable, that is the variable associated with the disease status. We simulate the omitted variable O and the genotype G to be independent in the population.

2.1.1 | Setting 1

We first examine models with one genetic variant. We simulate the genetic variant from Bernoulli(0.1) and an omitted variable O from Normal $(0, \sigma^2)$. We set $\sigma = 1, 2$ and next simulate the disease status according to the full disease risk model

$$\mathrm{logit}\{pr_B(D = d\,|\,G, O)\} = \beta_0 + \beta_G \times G + \beta_0 \times O, \quad (1)$$

where we let $\beta_0 = -1, -5$; $\beta_G = \log(1), \log(1.5), \log(2), \log(2.5), \log(3), \log(5), \log(8)$ and $\beta_0 = \log(1), \log(1.5), \log(2), \log(3), \log(5), \log(8)$, across various settings.

We generate 5,000 samples of 3,000/10,000 cases and 3,000/10,000 controls using a retrospective/case-control design.

We next estimate the parameters based on the reduced (and hence misspecified) logistic regression model

$$\log\mathrm{it}\{pr_{\Gamma}(D = d\,|\,G)\} = \gamma_0 + \gamma_{\mathrm{G}} \times G. \quad (2)$$

Shown in Table S1 are disease prevalence rate in the population, the difference $(\hat{\gamma}_G - \beta_G)$ that we define to be bias, variance, MSE, and FDR when $\beta_G = 0$.

The estimates in this setting are nearly unbiased with FDR being close to its nominal level. Shown in Table 1 is the setting when $\beta_G = \log(1.5)$. Here biases become more pronounced, leading to a reduced power for detecting an effect. For example, when

TABLE 1 Bias ($\hat{\gamma}_G - \beta_G$), variance and mean square error (MSE) of genetic effect estimates obtain using reduced model (2) when the data are simulated using full model (1)

(β_0, σ)	pr_B ($D = 1$)	Bias	Variance	MSE	Power
$\beta_O = \log(1)$					
(−1,1)	0.28	0.008	0.007	0.007	1
(−1,2)	0.28	0.008	0.007	0.007	1
(−5,1)	0.007	−0.03	0.005	0.006	1
(−5,2)	0.007	−0.03	0.005	0.006	1
$\beta_O = \log(1.5) = 0.4055$					
(−1,1)	0.28	−0.009	0.007	0.007	1
(−1,2)	0.30	−0.05	0.007	0.01	1
(−5,1)	0.008	−0.01	0.006	0.006	1
(−5,2)	0.01	−0.03	0.006	0.006	1
$\beta_O = \log(2) = 0.6931$					
(−1,1)	0.30	−0.03	0.008	0.009	0.99
(−1,2)	0.33	−0.10	0.007	0.02	0.95
(−5,1)	0.02	−0.03	0.006	0.007	1
(−5,2)	0.03	−0.03	0.006	0.007	1
$\beta_O = \log(2.5) = 0.9163$					
(−1,1)	0.31	−0.06	0.007	0.01	0.98
(−1,2)	0.35	−0.15	0.008	0.03	0.86
(−5,1)	0.01	−0.02	0.006	0.006	1
(−5,2)	0.03	−0.06	0.006	0.01	0.99
$\beta_O = \log(3) = 1.0986$					
(−1,1)	0.32	−0.07	0.007	0.01	0.98
(−1,2)	0.37	−0.18	0.007	0.04	0.76
(−5,1)	0.01	−0.01	0.006	0.006	1
(−5,2)	0.04	−0.11	0.006	0.02	0.96
$\beta_O = \log(5) = 1.6094$					
(−1,1)	0.34	−0.13	0.007	0.02	0.92
(−1,2)	0.40	−0.23	0.007	0.06	0.54
(−5,1)	0.02	−0.04	0.006	0.007	1
(−5,2)	0.09	−0.19	0.007	0.04	0.73
$\beta_O = \log(8) = 2.0794$					
(−1,1)	0.36	−0.17	0.007	0.03	0.80
(−1,2)	0.42	−0.27	0.007	0.08	0.37
(−5,1)	0.04	−0.10	0.007	0.02	0.97
(−5,2)	0.14	−0.25	0.007	0.07	0.48

Note: Shown is also probability of the disease in the population, that is pr_B ($D = 1$), and false discovery rate (FDR). The genotype is simulated to be Bernoulli(0.1), the omitted variable is simulated from Normal(0, σ^2). We simulated the disease status from model (1) with parameters $\beta_0 = -1, -5$; $\beta_G = \log(1.5)$, $\beta_O = \log(1), \log(1.5), \log(2), \log(2.5), \log(3), \log(5), \log(8)$. The results are based on 5,000 datasets of 3,000 cases and 3,000 controls.

TABLE 2 Bias ($\hat{\gamma}_G - \beta_G$), variance and mean square error (MSE) of genetic effect estimates obtain using reduced model (2) when the data are simulated using full model (1)

(β_0, σ)	pr_B ($D = 1$)	Bias	Variance	RMSE	Power
$\beta_O = \log(1) = 0$					
(−1,1)	0.32	−0.008	0.01	0.01	1
(−1,2)	0.32	−0.008	0.01	0.01	1
(−5,1)	0.01	−0.004	0.005	0.005	1
(−5,2)	0.01	−0.004	0.005	0.005	1
$\beta_O = \log(1.5) = 0.4055$					
(−1,1)	0.32	−0.07	0.01	0.02	1
(−1,2)	0.34	−0.25	0.01	0.07	1
(−5,1)	0.01	−0.02	0.005	0.005	1
(−5,2)	0.02	−0.06	0.005	0.009	1
$\beta_O = \log(2) = 0.6931$					
(−1,1)	0.33	−0.20	0.01	0.05	1
(−1,2)	0.36	−0.53	0.001	0.29	1
(−5,1)	0.01	−0.05	0.005	0.007	1
(−5,2)	0.02	−0.23	0.005	0.06	1
$\beta_O = \log(2.5) = 0.9163$					
(−1,1)	0.34	−0.31	0.01	0.10	1
(−1,2)	0.38	−0.74	0.009	0.56	1
(−5,1)	0.016	−0.07	0.005	0.01	1
(−5,2)	0.04	−0.45	0.005	0.21	1
$\beta_O = \log(3) = 1.0986$					
(−1,1)	0.35	−0.40	0.01	0.17	1
(−1,2)	0.39	−0.87	0.01	0.77	1
(−5,1)	0.019	−0.12	0.005	0.02	1
(−5,2)	0.05	−0.64	0.006	0.41	1
$\beta_O = \log(5) = 1.6094$					
(−1,1)	0.37	−0.64	0.009	0.42	1
(−1,2)	0.41	−1.17	0.008	1.37	1
(−5,1)	0.03	−0.34	0.005	0.12	1
(−5,2)	0.10	−1.04	0.006	1.09	1
$\beta_O = \log(8) = 2.0794$					
(−1,1)	0.39	−0.83	0.009	0.70	1
(−1,2)	0.43	−1.34	0.008	1.81	1
(−5,1)	0.05	−0.59	0.006	0.34	1
(−5,2)	0.15	−1.28	0.006	1.63	1

Note: Shown is also probability of the disease in the population, that is pr_B (D=1), and false discovery rate (FDR). The genotype is simulated to be Bernoulli(0.1), the omitted variable is simulated from Normal(0, σ^2). We simulated the disease status from model (1) with parameters $\beta_0 = -1, -5$; $\beta_G =$ log(8), $\beta_O =$ log(1), log(1.5), log(2), log(2.5), log(3), log(5), log(8). The results are based on 5,000 datasets of 3,000 cases and 3,000 controls.

$\beta_O = \log(3) = 1.0986, \beta_0 = -1, \sigma = 2,$ the difference ($\hat{\gamma}_G - \beta_G$) is 0.18, while the power to detect the genetic effect is 0.76. The frequency of the disease in the population is 0.37. Shown in Table 2 and Table S2 are the settings when $\beta_G =$ log(2), log(2.5), log(3), log(5), log(8). Biases increase with the coefficient β_G, however, because of its direction, the bias does not reduce power to detect the effect in this setting. As illustrated in Table S3 the biases noted in samples with 3,000 cases and 3,000 controls persist in samples with 10,000 cases and 10,000 controls.

2.1.2 | Setting 2

We now examine a setting with many genetic variables and one omitted environmental variable. The goal of this simulation is to see if the relative order of the genetic variables is estimated correctly. We simulate G_1 ... $G_{M/2}$ from Bernoulli(0.1) and $G_{M/2+1}$ ... G_M from Bernoulli (0.25), and O from Normal(0, σ^2), where $\sigma = 1, 2$.

Moreover, we simulate the disease status according to the risk model:

$$\log\mathrm{it}\{pr_B(D = 1 \mid G, O)\} = \beta_0 + \beta_{G_1} \times G_1 + \cdots + \beta_{G_M} \times G_M + \beta_0 \times O,$$

where we let $\beta_0 = -1; -5$; $M = 10$; $\beta_G \sim \mathrm{Normal}(\mu_G, \sigma_G^2)$, with $\mu_G = \log(1)$, $\log(2)$, $\log(3)$, $\log(5)$, $\log(8)$; $\sigma_G^2 = \log(1.5)$ and $\beta_O = \log(1)$, $\log(1.5)$, $\log(2)$, $\log(2.5)$, $\log(3)$, $\log(5)$, $\log(8)$. We next estimate the parameters based on the misspecified logistic regression model

$$\log\mathrm{it}\{pr_\Gamma(D = d \mid G)\} = \gamma_0 + \gamma_{G_1} \times G_1 + \cdots + \gamma_{G_M} \times G_M.$$

We would like to assess if the order of the genetic effect estimates is preserved. In many settings, it is more meaningful to estimate the relative order of the coefficients correctly, rather than the actual coefficients themselves. We define the order of the genetic effect by (a) absolute value of the coefficient; (b) *p*-value for the coefficient. The former is based just on the value of the coefficient estimate, while the latter is based on the *p*-value.

Shown in Table 3 and Table S4 are the results based on 5,000 samples of 3,000 cases and 3,000 controls. Shown in Table 3, the proportions of the genetic variants for which the ranks are preserved. As illustrated in Table S4, the proportions of the genetic variants for which the ranks are the same are very close to 1 when β_O is small.

In summary, we conclude that in the context of the genetic association studies the issue of bias due to omitting variables needs to receive more attention because it can be pronounced, in either direction and can distort false positive rate and power to detect an effect.

TABLE 3 Proportions of genetic variants that received the same rank based on the full and reduced genetic models across all variants (ALL), top 10% and top 20%

		$\hat{OR}$ Ranks based on			Ranks based on *p*-values		
(β_0, σ)	**pr_B ($D = d$)**	**ALL**	**TOP 10%**	**TOP 20%**	**ALL**	**TOP 10%**	**TOP 20%**
$\mu_G = \log(1)$							
(−1,1)	0.33	0.89	1	0.99	0.74	0.89	0.89
(−1,2)	0.37	0.77	0.98	0.93	0.58	0.81	0.75
(−5,1)	0.01	0.89	1	0.99	0.75	0.89	0.87
(−5,2)	0.05	0.79	0.98	0.95	0.57	0.69	0.71
$\mu_G = \log(2) = 0.693$							
(−1,1)	0.56	0.84	0.87	0.9	0.85	0.74	0.79
(−1,2)	0.55	0.71	0.73	0.77	0.70	0.53	0.61
(−5,1)	0.07	0.86	0.87	0.90	0.86	0.75	0.80
(−5,2)	0.12	0.74	0.73	0.79	0.73	0.53	0.62
$\mu_G = \log(3) = 1.099$							
(−1,1)	0.66	0.82	0.85	0.88	0.87	0.92	0.95
(−1,2)	0.63	0.70	0.73	0.77	0.76	0.74	0.83
(−5,1)	0.13	0.85	0.84	0.88	0.92	0.77	0.88
(−5,2)	0.18	0.73	0.73	0.77	0.80	0.62	0.79
$\mu_G = \log(5) = 1.609$							
(−1,1)	0.75	0.81	0.83	0.87	0.84	0.91	0.81
(−1,2)	0.71	0.69	0.71	0.76	0.72	0.84	0.74
(−5,1)	0.24	0.84	0.86	0.89	0.91	0.94	0.96
(−5,2)	0.28	0.71	0.72	0.77	0.80	0.85	0.88
$\mu_G = \log(8) = 2.079$							
(−1,1)	0.79	0.79	0.79	0.84	0.86	0.87	0.88
(−1,2)	0.76	0.67	0.66	0.72	0.74	0.78	0.68
(−5,1)	0.35	0.83	0.86	0.89	0.88	0.96	0.88
(−5,2)	0.37	0.71	0.71	0.77	0.77	0.89	0.77

Note: We simulated 5,000 datasets with 3,000 cases and 3,000 controls. We simulated 10 genetic variants from Bernoulli(0.1) and disease status from the full model with coefficients $\beta_O = \log(3)$ and $\mu_G = \log(1)$, $\log(2)$, $\log(3)$, $\log(5)$, $\log(8)$.

2.2 | Estimates of reduced versus full models

Suppose we obtained estimates of the genetic effects from a case-control study that omits a variable, that is, the coefficients in the reduced model (2). The risk of the disease is, however, determined by both the genetic effects G and the omitted variable O, that is the full model (1). We note that

$$\frac{pr(D = 1 \mid G, O)}{pr(D = 0 \mid G, O)} = \frac{f(O \mid G, D = 1) \times pr(D = 1 \mid G)}{f(O \mid G, D = 0) \times pr(D = 0 \mid G)}. \quad (3)$$

Hence

$$\log\mathrm{it}\{pr_\Gamma(D = 1 \mid G = g)\} = \beta_0 + \beta_G \times G + \beta_O \times O + \log\left\{\frac{f(O \mid D = 0, G = g)}{f(O \mid D = 1, G = g)}\right\}. \quad (4)$$

If the ratio $\frac{f(O \mid D=0, G=g)}{f(O \mid D=1, G=g)}$ does not depend on g, that is a constant of g, then the estimate of γ_G is consistent as the estimate of β_G. Hence if the omitted variable and the genotype are independent conditionally on the disease status D, then the reduced model yields consistent estimates of the genetic effects.

Bias recovery from $[O|D=d,\ G=g]$ Interestingly, when $[O|D=d,\ G=g] = \text{Normal}\ (\mu_0+\mu_g\times g+\mu_d\times d, \sigma^2)$, we derive that

$$\begin{aligned}\operatorname{logit}\{pr_{\Gamma}(D=1 \mid G=g)\} &= \beta_0+\beta_G\times G+\beta_O\times O \\ &\quad -\frac{\mu_d}{\sigma^2}\times O+\frac{\mu_d}{\sigma^2}\times\left(\mu_0+\mu_g\times g\right)+\frac{\mu_d^2}{2\times\sigma^2}.\end{aligned} \tag{5}$$

By Equation (5), we can derive $\beta_O = \frac{\mu_d}{\sigma^2}$ and $\gamma_G = \beta_G + \frac{\mu_d\mu_g}{\sigma^2}$. Therefore, the difference between γ_G and β_G is positive if $\mu_d\times\mu_g > 0$, while the difference between γ_G and β_G is negative if $\mu_d\times\mu_g < 0$. When the disease D is conditionally independent of the omitted variable O given genotype G, that is, $\beta_O = 0$, or equivalently, $\mu_d = 0$, then estimates of γ_G is also a consistent estimate of β_G. It is worth mentioning that in this setting we do not require independence between G and O.

Bias recovery from $[O]$ and $pr(D=1)$ When the auxiliary information of the omitted variable O is available in the literature and the rate of disease is known, we show that estimates of β_0, β_G, and β_O can be derived by solving a system of estimating equations. Here we assume that the omitted variable O follows a normal distribution Normal $(0, \sigma^2)$ and G follows the Bernoulli distribution; moreover, we assume that O and G are independent.

Based on the true model (1) and the fact that the rate of disease in the population, that is $pr\ (D=1)$, is known, we can obtain

$$E\left(\frac{e^{\beta_0+\beta_G\times G+\beta_O\times O}}{1+e^{\beta_0+\beta_G\times G+\beta_O\times O}}\right) = pr(D=1). \tag{6}$$

Suppose the estimated effect of the omitted variable O in the following reduced model:

$$\operatorname{logit}\{pr_{\mathrm{A}}(D=d \mid O)\} = \alpha_0+\alpha_O\times O. \tag{7}$$

has been reported in the literature. Under the full model (7), α_0 and α_O are the solutions to the expected score equations and thus we can derive

$$E\left\{O\left(\frac{e^{\beta_0+\beta_G\times G+\beta_O\times O}}{1+e^{\beta_0+\beta_G\times G+\beta_O\times O}} - \frac{e^{\alpha_0+\alpha_O\times O}}{1+e^{\alpha_0+\alpha_O\times O}}\right)\right\} = 0. \tag{8}$$

In a similar way, under the logistic regression model (2), γ_0 and γ_G are the solutions to the expected score equations and thus we can derive

$$E\left\{G\left(\frac{e^{\beta_0+\beta_G\times G+\beta_O\times O}}{1+e^{\beta_0+\beta_G\times G+\beta_O\times O}} - \frac{e^{\gamma_0+\gamma_G\times G}}{1+e^{\gamma_0+\gamma_G\times G}}\right)\right\} = 0. \tag{9}$$

Since α_0, α_O, and $pr(D{=}1)$ are known from the literature, we can calculate σ by solving $\int \frac{e^{\alpha_0+\alpha_O\times o}}{1+e^{\alpha_0+\alpha_O\times o}}\frac{1}{\sqrt{2\pi}}e^{\frac{-o^2}{2\sigma^2}}d\,o = pr(D=1)$. Based on the observed samples of G and D, we can derive consistent estimates for γ_G and $pr\ (G=1)$ and then an unbiased estimate for γ_0 can be obtained by solving $E\left(\frac{e^{\gamma_0+\gamma_G\times G}}{1+e^{\gamma_0+\gamma_G\times G}}\right) = pr(D=1)$. Applying numerical approximation to (6), (8), and (9), we can derive three estimating equations that only involve three unknown parameters β_0, β_G, and β_O. Consequently, we can derive estimates for β_0, β_G, and β_O by solving the three estimating equations.

2.3 | Simulation studies

The goal of the simulation studies is to assess bias in the estimates and the derivation (5) and the system of Equations (6), (8) and (9). We simulate the genetic variable G from Bernoulli(0.1).

2.3.1 | Setting 3

We are first interested to assess the Equation (5). Hence we simulate genotype from Bernoulli(0.1), then assume $\mu_0 = 0$, $\mu_g = \log(1.5)$, $\mu_d = -\log(1.5)$, $\log(1)$, $\log(1.5)$, $\sigma^2 = 1$, $\beta_0 = -1$, -3.5, $\beta_G = -\log(2.5)$, $-\log(1.5)$, $\log(1.5)$, $\log(2.5)$ and $\beta_O = \frac{\mu_d}{\sigma^2}$. Next we generate the disease status according to model (4) for 5,000 datasets with 3,000 cases and 3,000 controls. Shown in Table 4 and Table S5 are biases estimated based on (5), empirical bias, variance, MSE, and power. The results suggest that the empirical biases are similar to the biases obtained through (5).

2.3.2 | Setting 4

To evaluate the robustness of the proposed method when the model (6) is misspecified, we consider adding a quadratic term $\eta\times g^2$, where the genetic variant is generated from Bernoulli(0.1) and the omitted variable is generated from Normal(0, 1). We simulated 5,000 datasets with 3,000 cases and 3,000 controls. Moreover, we let $\mu_0 = 0, \mu_g = \mu_d = \log(1.5), \beta_G = \log(2.5), \beta_O = \frac{\mu_d}{\sigma^2}, \beta_0 = -1, -3.5$ and

$\eta = -\log(2.5)$, $-\log(1.5)$, $\log(1.5)$, $\log(2.5)$. As shown in Table 6, the estimates are substantially biased, hence the derived approximation is not robust to this degree of misspecification.

2.3.3 | Setting 5

We now assess the solution according to the system of Equations (6), (8), and (9). We simulate the genetic variant from Bernoulli(0.1) and the omitted variable from Normal(0, 1). And next we generate the disease status according to model (1) with coefficients $\beta_0 = -1, -5$; $\beta_G = \log(2.5), \log(3), \log(5), \log(8)$, $\beta_O = \log(5), \log(8)$ for 5,000 datasets with 3,000 cases and 3,000 controls. Results shown in Table 5 demonstrate that the numerical solution to the system of Equations (6), (8), and (9) is nearly unbiased.

2.3.4 | Setting 6

To evaluate robustness of the system of Equations (6), (8), and (9) to misspecification of the risk model, we simulate the data by adding an interaction term $\eta \times G \times O$ in model (1), while we recover the bias based on the system of Equations (6), (8), and (9). We generate the genetic variant from Bernoulli(0.1) and the omitted variable from Normal(0, 1). Moreover, we let $\beta_0 = -1, -5$, $\beta_O = \beta_G = \log(8)$, and $\eta = -\log(3)$, $-\log(1.5)$, $\log(1.5)$, $\log(3)$. We

TABLE 4 Bias approximation obtained using (5), that is $\frac{\mu_d \mu_g}{\sigma^2}$, rate of the disease in the population $pr_B\ (D = d)$, bias, variance, and mean squared error (MSE) of the estimates obtained from the reduced model

β_G	μ_d	$\frac{\mu_d \mu_g}{\sigma^2}$	$pr_B\ (D = 1)$	Bias	Variance	MSE	Power
$\beta_0 = -3.5$							
−log(2.5)	−log (1.5)	−0.16	0.03	−0.19	0.003	0.04	1
−log(2.5)	log (1)	0	0.03	0.01	0.003	0.003	1
−log(2.5)	log (1.5)	0.16	0.03	0.17	0.003	0.03	1
log(2.5)	−log (1.5)	−0.16	0.04	−0.16	0.002	0.03	1
log(2.5)	log (1)	0	0.03	0.01	0.002	0.002	1
log(2.5)	log (1.5)	0.16	0.04	0.17	0.002	0.03	1
$\beta_0 = -1$							
−log(2.5)	−log (1.5)	−0.16	0.27	−0.17	0.003	0.03	1
−log(2.5)	log (1)	0	0.25	0.01	0.003	0.003	1
−log(2.5)	log (1.5)	0.16	0.27	0.18	0.003	0.03	1
log(2.5)	−log (1.5)	−0.16	0.30	−0.16	0.002	0.03	1
log(2.5)	log (1)	0	0.29	0.002	0.002	0.002	1
log(2.5)	log (1.5)	0.16	0.31	0.17	0.002	0.03	1
$\beta_0 = -3.5$							
−log(1.5)	−log (1.5)	−0.16	0.03	−0.13	0.002	0.02	1
−log(1.5)	log (1)	0	0.03	0.02	0.002	0.003	1
−log(1.5)	log (1.5)	0.16	0.03	0.18	0.002	0.04	1
log(1.5)	−log (1.5)	−0.16	0.03	−0.15	0.002	0.02	1
log(1.5)	log (1)	0	0.03	0.01	0.002	0.002	1
log(1.5)	log (1.5)	0.16	0.03	0.17	0.002	0.03	1
$\beta_0 = -1$							
−log(1.5)	−log (1.5)	−0.16	0.27	−0.17	0.002	0.03	1
−log(1.5)	log (1)	0	0.26	0.01	0.002	0.003	1
−log(1.5)	log (1.5)	0.16	0.28	0.18	0.002	0.03	1
log(1.5)	−log (1.5)	−0.16	0.29	−0.17	0.002	0.03	1
log(1.5)	log (1)	0	0.28	0.005	0.002	0.002	1
log(1.5)	log (1.5)	0.16	0.30	0.17	0.002	0.031	1

Note: We simulated 5,000 datasets with 3,000 cases and 3,000 controls. We simulated genotype from Bernoulli(0.1), then assumed $\mu_0 = 0$, $\mu_g = \log(1.5)$, $\mu_d = -\log(1.5), \log(1.5)$, $\sigma^2 = 1$, $\beta_0 = -1, -3.5$, $\beta_G = -\log(2.5), -\log(1.5), \log(1.5), \log(2.5)$.

simulated 5,000 datasets with 3,000 cases and 3,000 controls. Results shown in Table 7 illustrate the recovery of bias is robust to this misspecification when the disease is common and when the effect size of the interaction G×O is smaller.

2.3.5 | Setting 7

We also examine the robustness of the recovery through the system of Equations (6), (8), and (9) in the setting where G and O are simulated to be correlated, but the analyses assume independence. We generated G from Bernoulli(0.1) and O from Normal $(c \times G, 1)$, where $c = 0.1$, 0.25, 0.5, 1. We generated the disease status the full model with $\beta_0 = -1$, -5 and $\beta_O = \beta_G = \log(8)$. We simulated 5,000 datasets with 3,000 cases and 3,000 controls. As shown in Table 8, the recovery of bias is robust to smaller degrees of correlation and smaller effect sizes.

TABLE 5 Bias, variance, and mean squared error (MSE) for the genetic effect estimates corrected based on the system of Equations (6), (8), and (9)

(β_0, β_O)	$pr_B\ (D = d)$	Bias	Variance	MSE
$\beta_G = \log(2.5)$				
(−1,log(5))	0.35	−0.006	0.02	0.02
(−5,log(5))	0.02	0.005	0.01	0.01
(−1,log(8))	0.37	0.009	0.03	0.03
(−5,log(8))	0.04	0.02	0.02	0.02
$\beta_G = \log(3)$				
(−1,log(5))	0.35	0.002	0.02	0.02
(−5,log(5))	0.02	0.007	0.01	0.01
(−1,log(8))	0.37	−0.002	0.03	0.03
(−5,log(8))	0.04	0.004	0.02	0.02
$\beta_G = \log(5)$				
(−1,log(5))	0.36	0.01	0.02	0.02
(−5,log(5))	0.03	0.006	0.02	0.02
(−1,log(8))	0.38	−0.02	0.03	0.03
(−5,log(8))	0.04	−0.003	0.02	0.02
$\beta_G = \log(8)$				
(−1,log(5))	0.37	0.002	0.03	0.03
(−5,log(5))	0.03	0.03	0.02	0.02
(−1,log(8))	0.39	−0.008	0.04	0.04
(−5,log(8))	0.05	0.02	0.03	0.03

Note: We simulated 5,000 datasets of 3,000 cases and 3,000 controls. The genetic variant is simulated Bernoulli(0.10), the omitted variable is simulated from Normal(0, 1) and the disease status is simulated based on model (2) with coefficients $\beta_0 = -1, -5$; $\beta_G = \log(2.5), \log(3), \log(5), \log(8)$, $\beta_O = \log(5), \log(8)$.

2.4 | Alzheimer's disease study

We are interested to assess what happens to the genetic effect estimates when a continuous variable is omitted from the model, that is how well (5) informs bias and if the system of Equations (6), (8), and (9) is capable to restore the genetic estimates. We hence consider two datasets. The Alzheimer's Disease Neuroimaging Initiative (ADNI) data set includes more extensive evaluations on a smaller subset of cases and controls. Next, we consider a larger data set generated by the Alzheimer's Disease Genetics Consortium (ADGC) where only a brief set of nongenetic variables is available on a large number of cases and controls. We hence assess how knowledge from the literature or from the ADNI data can be applied to inform how the genetic effect estimates obtained using ADGC change with the omission of continuous variables.

2.4.1 | Alzheimer's disease neuroimaging initiative

The set consists of 423 cases and 192 controls. We mapped the genetic variants to a set serving amyloid and tau proteins that are relevant to AD pathophysiology based on the Genecards database (https://www.genecards.org/). After preprocessing, the set contains 2,438 SNPs. The average (SD) age of cases is 74.29 (7.4), and 75.41 (4.91) in controls, $p = .058$. A total of 262 (61.9%) of cases are ApoE ε4 carriers and 49 (25.5%) of controls are ApoE ε4 carriers, $p < .001$. Table S7A–S7M further describes the sets of cases and controls and Web-based Supporting Information Materials Section B provides extended details on the analyses of the ADNI data set.

To assess what happens to the genetic effect estimates when a continuous variable is omitted from the model, we consider several possible models and the logistic regression results, including coefficient estimates (log (OR)), standard errors (*SE*) and *p*-values are reported in Table S8. We first consider a full Model 1 with age, sex, education, ApoE ε4 status, MMSE and a reduced model that omits MMSE (model 1A) and that omits ApoE ε4 status (model 1B). We next considered a full model 2 where we added ratio of hippocampus volume to whole-brain volume to model 1 with the corresponding reduced model that omits the ratio of brain volumes. We observed that the difference in log(OR) estimates between the reduced and full models were on the order of ≥1*SE. For example, log(OR) for ApoE ε4 status changed from 1.27 ($SE = 0.25$) to 1.62 ($SE = 0.20$) in full model 1 versus reduced model 1A; and from 1.005 ($SE = 0.268$) to 1.27 ($SE = 0.25$) in the full model 2 versus reduced model 2A.

TABLE 6 Analyses of robustness of model (6) when the true model includes an additional quadratic term $\eta \times G^2$, but the term is ignored in the analyses

β_0	$\frac{\mu_d \mu_g}{\sigma^2}$	$pr_B\ (D = d)$	Bias	Variance	MSE	Power
$\eta = -\log(2.5)$						
−1	0.1644	0.2888	−0.7455	0.0022	0.5580	0.9588
−3.5	0.1644	0.0324	−0.7476	0.0018	0.5607	0.9664
$\eta = -\log(1.5)$						
−1	0.1644	0.3008	−0.2331	0.0021	0.0564	1
−3.5	0.1644	0.0347	−0.2524	0.0016	0.0653	1
$\eta = \log(1.5)$						
−1	0.1644	0.3204	0.5669	0.0026	0.0324	1
−3.5	0.1644	0.0412	0.5612	0.0015	0.0317	1
$\eta = \log(2.5)$						
−1	0.1644	0.3313	1.0812	0.0032	1.1723	1
−3.5	0.1644	0.0482	1.0858	0.0016	1.1807	1

Note: The approximation to the bias, rate of the disease in the population $pr_B\ (D = d)$, empirical bias, variance, and mean squared error (MSE) of the estimates obtained from the reduced model. We simulated 5,000 datasets with 3,000 cases and 3,000 controls. We simulated genotype from Bernoulli(0.1) and the omitted variable from Normal(0, 1), then assumed and $\eta = -\log(2.5), -\log(1.5), \log(1.5), \log(2.5)$.

TABLE 7 Evaluation of robustness of model (1) when the true model includes an additional interaction term $\eta \times \mathrm{G} \times \mathrm{O}$

β_0	$pr_B\ (D = d)$	Bias	Variance	MSE
$\eta = -\log(3)$				
−1	0.3902	0.2970	0.0399	0.1281
−5	0.0391	−1.0881	0.0138	1.1978
$\eta = -\log(1.5)$				
−1	0.3867	0.0898	0.0398	0.0479
−5	0.0430	−0.4126	0.0187	0.1889
$\eta = \log(1.5)$				
−1	0.3830	−0.1142	0.0483	0.0614
−5	0.0485	0.3188	0.0419	0.1436
$\eta = \log(3)$				
−1	0.3809	−0.2830	0.0544	0.1345
−5	0.0527	0.8983	0.0536	0.8605

Note: Shown are bias, variance, and mean squared error (MSE) of the estimates for the genetic effect corrected based on the system of Equations (6), (8), and (9). We simulated 5,000 datasets with 3,000 cases and 3,000 controls. We simulated genotype from Bernoulli(0.1) and the omitted variable from Normal(0, 1), then assumed $\beta_O = \beta_G = \log(8)$, and $\eta = -\log(3), -\log(1.5), \log(1.5), \log(3)$.

We next assumed that the full include age, gender, education ApoE ε4 status, MMSE, and the ratio between hippocampus volume and whole-brain volume, plus each of the genetic variants (Model 3). The reduced model 3A omits the ratio between brain volumes. On average, we observed that the difference in the log(OR) estimates of SNPs obtained in full model 3 versus 3A is 0.006, with 25th percentile −0.001 and 75th percentile that is 0.005, minimum of −0.35 and maximum of 0.18.

TABLE 8 Analyses of robustness to the correlation between G and O. We simulated G from Bernoulli(0.1) and O from Normal $(c \times G, 1)$, where $c = 0.1$, 0.25, 0.5, 1

β_0	$pr_B\ (D = d)$	Bias	Variance	MSE
$c = 0.1$				
−1	0.3876	−0.1879	0.0205	0.0558
−5	0.0476	−0.1052	0.0086	0.0197
$c = 0.25$				
−1	0.3915	−0.0551	0.0199	0.0230
−5	0.0504	0.0726	0.0120	0.0173
$c = 0.5$				
−1	0.3974	0.0555	0.0129	0.0160
−5	0.0559	0.3567	0.0179	0.1452
$c = 1$				
−1	0.4066	−0.1051	0.0068	0.0178
−5	0.0694	0.7819	0.0432	0.6546

Note: Shown are bias, variance, and mean squared error (MSE) of the estimates for the genetic effect estimates corrected based on the system of Equations (6), (8), and (9). We simulated 5,000 datasets with 3,000 cases and 3,000 controls. We assumed $\beta_0 = -1, -5$ and $\beta_O = \beta_G = \log(8)$.

We observed in the following about how the SNPs rank in the full model 3 and reduced model 3A. Among the top 10 significant SNPs (ranked by *p*-value), 80% of the SNPs are the same in the full and reduced models, among the top 30 significant SNPs, 56.67% of the SNPs are the same and among the top 50 significant SNPs, 58% of the SNPs are the same. Hence overall, the conclusion about what SNPs should be carried to the validation set would be different based on these two models.

We also note that for all the models the distribution of *p*-values across all SNPs did not differ significantly from Uniform(0,1), that is *p*-values for Kolmogorov-Smirnov test are >.05.

2.4.2 | Alzheimer's disease genetics consortium

The set consists of 2,794 cases and 667 controls (Set 1), where subsets contained data on age, sex, education, ApoE $\varepsilon4$ status (Set 2). We mapped the genetic variants to a set serving innate immune system that are relevant to AD pathophysiology (Lobach, Kim, Alekseyenko, Lobach, & Zhang, 2019). After processing, the set contains 157 SNPs. The average (SD) age of cases is 70.78 (8.82), and 75.19 (8.27) in controls, $p < .001$. The average (SD) education of cases is 14.08 (3.38), and 15.93 (2.72) in controls, $p < .001$, 1,005 (48.9%) of cases are men, 109 (32.8%) of controls are men, $p < .001$. A total of 1,327 (64.6%) of cases are ApoE $\varepsilon4$ carriers and 96 (28.9%) of controls are ApoE $\varepsilon4$, $p < .001$. The data set and analyses are described in extensive detail in Web-based Supporting Information Materials Section C.

We first assessed estimates in the full and reduced models based on a subset of data that includes age, sex, education, ApoE $\varepsilon4$. We observed that estimates of SNPs differed between the full (age, sex, education, ApoE $\varepsilon4$, SNP) and reduced models (omits age) by on average 0.01, 25th percentile = −0.02, 75th percentile = 0.04, minimum of 0.17 and maximum of 0.58.

We also note that for all the models the distribution of *p*-values across all SNPs did not differ significantly from Uniform(0,1), that is *p*-values for Kolmogorov-Smirnov test are >.05.

We are next interested in asses the degree and directionality to which estimates of ApoE $\varepsilon4$ status change with the omission of age, MMSE, education, hippocampal volume and the ratio of the hippocampal volume to the whole brain volume. We, therefore, consider the set of 2,794 cases and 667 controls. We first estimate $\gamma_{\varepsilon4}$ from a univariable model to be 0.16 ($SE = 0.01$), $p < .001$. We next learn the conditional distributions $[O|D = d, \varepsilon4]$ of each of the omitted variables from the ADNI data set, where we define the set of cases to be the set with diagnosis dementia and the set of controls to be the set with a diagnosis of cognitively normal. Then we apply the relationship (6) to estimate the difference in the estimates due to omitting the variable as $\frac{\mu_d \mu_g}{\sigma^2}$. As the result, we estimate that omission of age decreases the log(OR) for ApoE $\varepsilon4$ status by 0.10, omission of MMSE increases the estimate by 0.10, omission of education increases the estimate by 0.04, omission of hippocampal brain volume increases the estimate by 0.06, and omission of the ratio between hippocampal brain volume and whole-brain volume increases the estimate by 0.06.

We next assessed how the coefficient for ApoE $\varepsilon4$ status changes when MMSE is omitted from the model using the system of Equations (6), (8), and (9). From the literature, we assumed that MMSE is distributed normally with mean 27 and standard deviation of 1.8; frequency of the disease in the population that is 10% and OR for MMSE that is 0.8 (95% CI: 0.55–1.1). In the reduced model the log(OR) for ApoE $\varepsilon4$ status is 1.5 ($SE = 0.13$), $p < .001$. Using the system of Equations (6), (8), and (9) we arrived at the log(OR) estimate that varied between 1.45 and 1.79 for various settings of the initial values that we considered.

3 | DISCUSSION

In the genetic association studies, we interested to accurately estimate either the parameters or the order of the magnitude of the parameters, because the estimates would determine our understanding about the underlying pathophysiologic mechanisms, risk prediction and can lead to the estimates of heritability, population attributable risk to the genetics, and so forth. Massive amounts of genetic data available in various databases can be utilized to estimate the genetic associations. Yet, the set of nongenetic variables is often brief.

We show that omitting a continuous variable associated with the disease status can result in substantial bias of parameter estimates in either direction. We derived two possible approaches to understanding the bias. The first is explicit and is based on knowing $[O|D = d, G = g]$. The second is numerical and requires knowing the estimates from a univariable model with the omitted variable (8) and knowing the rate of the disease in the population as well as the distribution of the omitted variable in the population. The two approaches that we developed differ in their assumptions. One assumes a Normal distribution for the conditional density of the omitted variable $[O|D = d, G = g]$, that is assumes that the distribution of the omitted variable is a mixture of normals. The second in the system of equations assume that the distribution of the omitted variable is normal and that the genetic variables and the omitted variable are independent.

Both of the approaches that we considered require knowing the set of variables in the full (true) model. In the analyses of Alzheimer's disease studies, we assumed various models to be the true (full) models and based on these assumptions assessed the directionality and magnitude of bias. Overall, the main contribution of our work is the justification that omitting a continuous variable from the logistic regression model can result in bias in either direction.

In some settings, it is of interest to correctly estimate the order of the magnitude of the genetic effects to be able to rank the genetic markers according to the strength of their association. In these settings, if the bias affects the estimates proportionally, then the bias would not change the ordering of the genetic effect estimates.

We found that if the genetic variable and the omitted variable are independent conditionally on the disease status, then omitting the variable does not result in bias of the genetic effects. This assumption is not equivalent to independence between the genotype and the omitted variable in the population.

The arguments that we've developed are based on the logistic link model and normality of the omitted variable. These derivations do not naturally extend to other link functions and other forms of the omitted variable.

Pirinen, Donnelly, and Spencer (1993) showed that for rare diseases inclusion of the key covariates can reduce power, while for common diseases inclusion of the key covariates can increase power. Our findings are similar in that the bias can either reduce or increase the magnitude of the effect. Specifically, if the omitted variable is normally distributed with $[OD = d, G = g] =$ Normal $(\mu_0 + \mu_g \times g + \mu_d \times d, \sigma^2)$ then the bias is a function of μ_g, μ_d, and σ^2. Based on this relationship we also see that a rare disease is not immune to the bias.

ACKNOWLEDGMENTS

Dr. Lobach is supported by 5R21AG043710-02.

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

Genotyping is performed by Alzheimer's Disease Genetics Consortium (ADGC), U01 AG032984, RC2 AG036528. The 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)

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRxResearch; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the Database of Genotypes and Phenotypes at ([https://www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgi?study_id=phs000372.v1.p1), reference number [phs000372.v1.p1] and at the Alzheimer's Disease Neuroimaging Initiative http://adni.loni.usc.edu/. Alzheimer's Disease Genetics Consortium (ADGC) Genome-Wide Analyses Association Study—NIA Alzheimer's Disease Centers Cohort, https://www.ncbi.nlm.nih.gov/projects/

gap/cgi-bin/study.cgi?study_id=phs000372.v1.p1 and Alzheimer's Disease Neuroimaging Initiative http://adni.loni.usc.edu.

ORCID

Iryna Lobach http://orcid.org/0000-0002-3317-7633

REFERENCES

Gail, M. H., Wieand, S., & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, *71*(3), 431–444.

Hauck, W. W., Neuhaus, J. M., Kalbfleisch, J. D., & Anderson, S. (1991). A consequence of omitted covariates when estimating odds ratios. *Journal of Clinical Epidemiology*, *44*(1), 77–81.

Kukull, W. A., Schellenberg, G. D., Bowen, J. D., McCormick, W. C., Yu, C. E., Teri, L., … Larson, E. B. (1996). Apolipoprotein E in Alzheimer's disease risk and case detection: A case-control study. *Journal of Clinical Epidemiology*, *49*(10), 1143–1148.

Lobach, I., Kim, I., Alekseyenko, A., Lobach, S., & Zhang, L. (2019). A simple approximation to bias in gene-environment interactions when a case might not be the case. *Frontiers in Genetics. Statistical Genetics and Methodology*, https://doi.org/10.3389/fgene.2019.00886

Neuhaus, J. M., & Jewell, N. P. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, *80*(4), 807–815.

Park, J. H., Gail, M. H., Weinberg, C. R., Carroll, R. J., Chung, C. C., Wang, Z., … Chatterjee, N. (2011). Distribution of allele frequencies and effect sizes and their interrealtionships for commen genetic susceptibility variants. *Proceedings of the National Academy of Sciences USA*, *208*(44), 18026–18031.

Pirinen, M., Donnelly, P., & Spencer, C. A. (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics*, *44*(8), 848–851.

Prentice, K. L., & Pyke, D. A. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, *66*(3), 403–411.

Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, *44*(4), 1049–1060.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Sheng Y, Huang C-Y, Lobach S, Zablotska L, Lobach I,. Genetic effect estimates in case-control studies when a continuous variable is omitted from the model. *Genetic Epidemiology*. 2020;44:261–271. https://doi.org/10.1002/gepi.22278