A simple approximation to bias in the genetic effect estimates when multiple disease states share a clinical diagnosis

Iryna Lobach[1*], Inyoung Kim[2], Alexander Alekseyenko[3], Siarhei Lobach[4], Li Zhang[5]

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

[2] Department of Statistics, Virginia Tech University, Blacksburg, VA, USA

[3] Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, USA

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

[5] Department of Medicine, University of California, San Francisco, San Francisco, USA

*Corresponding author:

Iryna Lobach, Ph.D.

Division of Biostatistics

Department of Epidemiology and Biostatistics

University of California, San Francisco

Email: Iryna.lobach@ucsf.edu

Phone: 415-476-6115

Running title: Bias in the genetic estimates when a case is not the case

ABSTRACT

Case-control genome-wide association (CC-GWAS) studies might provide valuable clues to the underlying pathophysiologic mechanisms of complex diseases, such as neurodegenerative disease, cancer. A commonly overlooked complication is that multiple distinct disease states might present with the same set of symptoms and hence share a clinical diagnosis. These disease states can only be distinguished in a biomarker evaluation that might not be feasible on the whole set of cases in the large number of samples that are typically needed for CC-GWAS. Instead, the biomarkers are measured on a subset of cases. Or an external reliability study estimates frequencies of the disease states of interest within the clinically diagnosed set of cases. These frequencies often vary by the genetic and/or non-genetic variables. We derive a simple approximation that relates the genetic effect estimates obtained in a logistic regression model with the clinical diagnosis as an outcome variable to the estimates in the relationship to the true disease state of interest. We performed simulation studies to assess accuracy of the approximation that we've derived. We next applied the derived approximation to the analysis of the genetic basis of innate immune system of Alzheimer's disease.

INTRODUCTION

Case-control genome-wide analyses scan (CC-GWAS) is a tool that is widely used to elucidate the genetic basis of complex diseases. A common complication is that multiple distinct disease states share the observed symptoms and hence the clinical diagnosis. Frequencies of the disease states within the clinical diagnosis often vary by the key variables. If the disease states have distinct genetic basses, the analyses with a clinical diagnosis as an outcome variable might be substantially biased (Carroll et al, 2006).

The specific example that motivated this study is the analyses of the genetic susceptibility to Alzheimer's disease (AD). The clinical diagnosis of AD is typically made based on a set of descriptive criteria and only a small subset of cases receives positron emission tomography (PET) to evaluate for amyloid positivity, what is a requirement for the *true*, or pathologically defined, AD. Recent biomarker studies (Salloway and Sperling, 2015) estimate that 36% of ApoE ε4 non-carriers and 6% of ApoE ε4 carriers diagnosed with AD do not have evidence for amyloid as measured by PET, hence do not qualify for the *true* AD diagnosis.

We are interested to examine the role of the genetic variants serving the innate immune system in susceptibility to AD, i.e. the AD symptoms underlined by the amyloid deposition. The usual analyses define the outcome variable in a regression analysis to be the clinical diagnosis. We, however, recognize heterogeneity of the clinical diagnosis where the underlying disease state separates the cases into a subset with amyloid-related AD, what is the disease state of interest; and non-amyloid-related AD, what is the nuisance disease state. We derive the theoretical approximation that provides a simple and general relationship between B and Γ estimates using Kullback-Leibler divergence (Kullback, 1959).

Our paper is organized as follows. First, in the Material and Methods section we present the setting, notation, and the proposed approximation for various models. Next, in the Simulation Experiments section we describe the empirical studies that are conducted to compare the resulting performance of the approximation that we derived relative to the average observed across many simulated datasets. We then compare the estimates in a practical setting of an Alzheimer's disease study that aims to investigate the genetic basis of innate immune system in the relationship to the AD symptoms underlined by amyloid pathology. We conclude our paper with brief Discussion.

MATERIALS AND METHODS

We define G to be the genotype of single nucleotide polymorphisms (SNPs) measured at multiple locations. Let X and Z be the environmental variables that might interact. We assume that the genotype is independent of the environment and follows Hardy-Weinberg equilibrium model $Q(g;\theta)$, where θ is the frequency of the minor allele.

We define D^{CL} be the observed clinical diagnosis that is inferred based on a set of descriptive criteria that characterize symptoms. Let D denote the true disease states, where $D = 1$ indicates the disease state of interest and $D = 1^*$ is the nuisance disease state. It might not be possible to measure D on the set of cases in a GWAS, instead D is available on a subset or frequencies of D within the clinically defined set of cases are reliably estimated in an external reliability study. We define the clinical-pathological diagnosis relationship using $\tau(X) = pr(D = 1|D^{CL} = 1, X)$, what is a frequency of the disease state of interest within the clinically diagnosed set and the frequency varies by X. In the context of AD study, $pr(D = 1^*|D^{CL} = 1, X) = 1 - \tau(X)$, $pr(D = 0|D^{CL} = 1, X) = 0$, $pr(D = 1^*|D^{CL} = 0, X) = pr(D = 1|D^{CL} = 0, X) = 0$ and $pr(D = 1^*|D^{CL} = 0, X) = pr(D = 0|D^{CL} = 0, X) = 1$. We define the probabilities of the clinical diagnosis in the population to be $\pi_{d^{CL}} = pr(D^{CL} = d^{cl})$. Similarly, we let frequencies of the true pathologic state in the population to be $\pi_d = pr(D = d)$.

For clarity of presentation we assume that genotype is binary to indicate presence of a minor allele, environmental variables X and Z are Bernoulli with frequencies η_X and η_Z, respectively. In the Appendix we discuss how to extend the approximation to the categorical and continuous variables.

Model 1. $\boldsymbol{\beta_G}$: We first consider a setting when only the genetic variable G is in the risk model, i.e. the true disease risk model is

$$\log\left\{\frac{pr_{B,A}(D=1|G)}{pr_{B,A}(D=0|G)}\right\} = \beta_0 + \beta_G \times G;$$

(1)

$$\log\left\{\frac{pr_{B,A}(D=1^*|G)}{pr_{B,A}(D=0|G)}\right\} = \alpha_0 + \alpha_G \times G; \qquad (2)$$

while the model used is the usual logistic regression model with the clinical diagnosis as an outcome variable, i.e.

$$\log\left\{\frac{pr_\Gamma(D^{CL}=1|G)}{pr_\Gamma(D^{CL}=0|G)}\right\} = \gamma_0 + \gamma_G \times G. \qquad (3)$$

Derivations provided in Appendix A1 show that

$$\gamma_0 \approx \log\{\exp(\beta_0) + \exp(\alpha_0)\};$$

(4a)

$$\gamma_G \approx \log\{\exp(\beta_0 + \beta_G) + \exp(\alpha_0 + \alpha_G)\} - \log\{\exp(\beta_0) + \exp(\alpha_0)\}$$

$$\approx \log\{\exp(\beta_0) + \exp(\alpha_0 + \alpha_G)\} - \log\{\exp(\beta_0) + \exp(\alpha_0)\} + \frac{\exp(\beta_0)}{\exp(\beta_0)+\exp(\alpha_0+\alpha_G)} \times \beta_G. \qquad (4b)$$

From (4a) and (4b), we derive that

$$\beta_0 \approx \log\{\exp(\gamma_0) - \exp(\alpha_0)\};$$

(4c) $\beta_G \approx \log\{\exp(\gamma_0 + \gamma_G) - \exp(\alpha_0 + \alpha_G)\} - log\{\exp(\gamma_0) - \exp(\alpha_0)\}.$

(4d)

Appendix A3 describes how to obtain β_0, β_G, α_0, α_G, assuming estimates of γ_0, γ_G are available from the usual logistic regression and reliable estimates of $\tau = \tau(1) \times pr(X=1) + \tau(0) \times pr(X=0)$ and π_1 are available in the literature.

Model 2. $\boldsymbol{\beta_G}$ and $\boldsymbol{\beta_X}$: We next consider a setting when the genetic variable G and an environmental variable X are in the risk model, i.e. the true disease risk model is

$$\log\left\{\frac{pr_{B,A}(D=1|G,X)}{pr_{B,A}(D=0|G,X)}\right\} = \beta_0 + \beta_G \times G + \beta_X \times X;$$

(5)

$$\log\left\{\frac{pr_{B,A}(D=1^*|G,X)}{pr_{B,A}(D=0|G,X)}\right\} = \alpha_0 + \alpha_G \times G + \alpha_X \times X;$$

(6)

while the model used is

$$\log\left\{\frac{pr_{\Gamma}(D^{CL}=1|G,X)}{pr_{\Gamma}(D^{CL}=0|G,X)}\right\} = \gamma_0 + \gamma_G \times G + \gamma_X \times X.$$

(7)

Derivations provided in Appendix A2 show that

$$\gamma_0 \approx \log\{\exp(\beta_0) + \exp(\alpha_0)\}\,;$$

(8a)

$$\gamma_G \approx 0.5 \times \textstyle\sum_x [\log\{\exp(\beta_0 + \beta_G + \beta_X \times x) + \exp(\alpha_0 + \alpha_G + \alpha_X \times x)\} - \log\{\exp(\beta_0 + \beta_X \times x) + \exp(\alpha_0 + \alpha_X \times x)\}]$$

$$\approx 0.5 \times \textstyle\sum_x [\log\{\exp(\beta_0 + \beta_X \times x) + \exp(\alpha_0 + \alpha_G + \alpha_X \times x)\} - \log\{\exp(\beta_0 + \beta_X \times x) + \exp(\alpha_0 + \alpha_X \times x)\}] + 0.5 \times \sum_x \frac{\exp(\beta_0 + \beta_X \times x)}{\exp(\beta_0 + \beta_X \times x) + \exp(\alpha_0 + \alpha_G + \alpha_X \times x)} \times \beta_G\,;$$

(8b)

$$\gamma_X \approx 0.5 \times \textstyle\sum_g [\log\{\exp(\beta_0 + \beta_X + \beta_G \times g) + \exp(\alpha_0 + \alpha_X + \alpha_G \times g)\} - \log\{\exp(\beta_0 + \beta_G \times g) + \exp(\alpha_0 + \alpha_G \times g)\}]$$

$$\approx 0.5 \times \sum_g [\log\{\exp(\beta_0 + \beta_G \times g) + \exp(\alpha_0 + \alpha_X + \alpha_G \times g)\} - \log\{\exp(\beta_0 + \beta_G \times g) + \exp(\alpha_0 + \alpha_G \times g)\}] \ + 0.5 \times \sum_x \frac{\exp(\beta_0+\beta_G\times g)}{\exp(\beta_0+\beta_G\times g)+\exp(\alpha_0+\alpha_X+\alpha_G\times g)} \times \beta_X.$$

(8c)

Model 3. $\boldsymbol{\beta_G}$, $\boldsymbol{\beta_X}$, $\boldsymbol{\beta_Z}$, and $\boldsymbol{\beta_{X\times Z}}$: A model with interaction between the environmental variables X and Z is discussed in Appendix.

Model 4. $\boldsymbol{\beta_{G_1}}, \boldsymbol{\beta_{G_2}}$ and $\boldsymbol{\beta_{G_1\times G_2}}$: A model with gene-gene interactions is discussed in Appendix.

Remarks:

1. Model 1, equation (4b). If $\beta_G = \alpha_G = 0$, then $\gamma_G = 0$.
2. Model 2, equation (8b). If $\beta_G = \alpha_G = 0$, then $\gamma_G = 0$.
3. Model 2, equation (8c). If $\beta_X = \alpha_X = 0$, then $\gamma_X = 0$.
4. Remarks 1-3 describe when the usual logistic regression models with the clinical diagnosis as an outcome variable correctly estimate the null effect.
5. The equations that we derived apply to several possible likelihood functions. For example, parameter estimates in Model 3 can be estimated based on the usual logistic regression model, i.e. the probability of the form $pr_\Gamma(D^{CL}|G,X,Z)$ or in a pseudolikelihood (Spinka et al, 2005; Lobach et al, 2018) $pr_\Gamma(D^{CL},G|X,Z,\delta = 1)$, where $,\delta = 1$ is an imaginary indicator of being selected into the study. All the derivations apply to both models.

SIMULATION STUDIES

False positive rate We first perform a series of simulation experiments to examine a false positive rate in the estimates of β_G when the data are simulated from model (1)-(2), but the parameter estimates are obtained from model (3). We define the false positive rate to be the fraction of p-values≤0.05 across 10,000 simulated datasets in the usual logistic regression analyses as an outcome variable, i.e. (3), when in fact $\beta_G = 0$. We simulate the data using model (1) with coefficients $\beta_0 = 0.5, \beta_G = 0, \alpha_G = \log(1) = 0\,, \log(1.5) = 0.41, \log\,(2) = 0.69$. We next estimate parameters using model (3). **Table 1** presents false positive rates in datasets with $n_0 = n_1 = 3{,}000; 10{,}000$. When the genetic effect is not associated with the clinical diagnosis, the false positive rate is nominal, i.e. is nearly 0.05. When α_G increases, the false positive rate gets inflated, e.g. when $\alpha_G = \log(1.5) = 0.41$, the false positive rate is 0.72. Increase in sample size did not result in decrease of the false positive rate.

<u>**Approximation vs. empirical estimates**</u> We next perform a series of simulation experiments to assess the magnitude of bias and the approximation to the relationships that we've derived. First, we estimate the bias empirically as the average across 500 simulated datasets where the data are simulated using the true model (1)-(2), (5)-(6), (A3)-(A4) based on coefficients B and A, but estimate the parameters Γ in the usual logistic regression model (3), (7) and (A5). We then compare these averages to the approximations that we've derived.

We simulate genotype (G), age (A), sex (S), ApoE $\epsilon 4$ status to be Bernoulli with frequencies $\theta_G, \theta_A, \theta_S, \theta_{\epsilon 4}$. In the context of previous notations, X is the ApoE $\epsilon 4$ status and Z is a set consisting of G, A, S. We then simulated the clinical diagnosis status D^{CL} according to the models (3), (7) and (A5) and the true disease states D according to model (1)-(2), (5)-(6), (A3)-(A4). In all simulations we let $\theta_G = 0.10, \theta_A = 0.50, \theta_S = 0.52, \theta_{\epsilon 4} = 0.07$.

Model 1 We fist simulate the data using model (1)-(2) and estimate parameters in the logistic model (3). We set $\beta_0 = 0.5$, $\beta_G = \log(1) = 0, \log(1.5) = 0.41, \log(2) = 0.69, \log(2.5) = 0.92$, log(3)=1.1, $\alpha_G = \log(1) = 0\,, \log(1.5) = 0.41, = 0.69$ and simulate datasets with 3,000 cases and 3,000 controls. **Table 2** presents empirical estimates of β_G and the approximation (4b). Across all values of β_G and α_G, the approximation (4b) is accurate relative to the empirical estimate.

Model 2 We next generate data using models (5)-(6) but estimate parameters using model (7). We let $\beta_0 = \alpha_0 = 0.5$, $\beta_G = \log(1) = 0\,, \log(1.5) = 0.41\,, \log(2) = 0.69\,, \log(2.5) = 0.92\,, \log(3) = 1.1, \beta_{\epsilon 4} = \alpha_{\epsilon 4} = \log(8)\,, \alpha_G = \log(1) = 0, \log(2) = 0.41\,, \log(3) = 0.69, \log(4) = 1.1$ and generate datasets with 3,000 cases and 3,000 controls. Approximations and the empirical estimates for γ_G shown in **Table 3** demonstrate that the approximation (8b) is accurate relative to the empirical estimates. The empirical estimate of $\gamma_{\epsilon 4}$ is 2.09, while the approximation is 2.08.

Model 3 We next simulate data using models (A3)-(A4) and estimate parameters using model (A5), with the approximation derived in (A6a-c).

Setting 1. We first consider a setting when the nuisance disease is not associated with the genotype ($\alpha_G = 0$) and when $\epsilon 4$ and $A \times \epsilon 4$ are not associated with the nuisance disease status ($\alpha_{\epsilon 4} = 0, \alpha_G = 0, \alpha_{A\times\epsilon 4} = 0$). We simulate the clinical diagnosis and disease states with coefficients $\beta_0 = -1, \beta_S = log(0.92) = -0.08, \beta_{\epsilon 4} = log(8) = 2.1, \beta_A = log(2) = 0.69,$

$\beta_G = log(1), log(1.5), log(2), log(2.5), log(3), \beta_{A\times\epsilon 4} = log(1), log(2), log(3), log(3),$

$\alpha_0 = -1, \alpha_S = log(0.92), \alpha_{\epsilon 4} = 0, \alpha_A = log(2), \alpha_G = 0, \alpha_{A\times\epsilon 4} = 0.$

Figure 1 presents biases in estimates of β_G (panel A), β_A (panel B), β_S (panel C), $\beta_{\epsilon 4}$ (panel D) and $\beta_{A\times\epsilon 4}$ (panel E) in studies with 3,000 cases and 3,000 controls; values of $\beta_{A\times\epsilon 4}$ are shown along the x-axis and values of β_G are indicated by color. **Figure 1** panels **A** and **D** show that bias in the estimates of β_G and $\beta_{\epsilon 4}$ can be substantial with largest bias of -0.06; panel **E** shows that bias in $\beta_{A\times\epsilon 4}$ is notable in this case ranging from 0.01 to -0.06; estimates of β_A and β_S are nearly unbiased consistent with the theoretical observations that the null effect in some settings can be estimated with no bias even in a misspecified model. We note that magnitude of bias in $\widehat{\beta_G}$ and $\widehat{\beta_{A\times\varepsilon 4}}$ increases as the true value of the coefficient increases.

Shown on **Figure 2** are the empirical bias (Emp) and the approximation (AX) of bias in β_G indicated by color with values of $\beta_{A\times\epsilon 4}$ along the x-axis and values of β_G along the panels. The difference between the Emp and AX starts at ≈ 0.6 when $\beta_G = 0.41$ and increases to ≈ 1.2 when $\beta_G = 1.1$. Bias of $\widehat{\beta_S}$ and $\widehat{\beta_A}$ is approximated to be <0.0001. Shown on **Figure 3** are Emp and AX of estimates of $\beta_{\epsilon 4}$, and **Figure 4** is presenting estimates of $\beta_{A\times\epsilon 4}$.

Setting 2. We next simulate datasets with 30,000 cases and 30,000 controls in the Setting 1. **Supplementary Figure 1** shows that biases in the estimates noted in Setting 1 persists for larger sample sizes.

Setting 3. We next consider a setting when with the nuisance disease the genotype is associated ($\alpha_G = log(1.5)$), $\alpha_{\epsilon 4}$ is associated ($\alpha_{\epsilon 4} = log(2)$), and no interaction $\alpha_{A\times\epsilon 4} = 0$. We next change the parameters for the nuisance state to be $\alpha_0 = -1, \alpha_S = log(0.92), \alpha_{\epsilon 4} = log(2), \alpha_A = log(2), \alpha_G = log(1.5), \alpha_{A\times\epsilon 4} = 0$ and all other parameters as in Setting 1. Shown on **Supplementary Figure 2** are biases in the estimates of the parameters of interest that reach -1.4 for $\widehat{\beta_G}$, are near -0.5 for $\widehat{\beta_{\epsilon 4}}$, and can reach -0.15 for $\widehat{\beta_{A\times\varepsilon 4}}$.

Setting 4. We next consider a Setting 1 but with more common disease of interest, i.e. $\beta_0 = 1.5$. **Supplementary Figure 3** is showing empirical bias in all estimates. The estimates can still be substantially biased.

Setting 5. We next consider a setting where the genetic variable is associated with the nuisance disease state ($\alpha_G = log(2)$) and there is significant $A \times \epsilon 4$ interaction ($\alpha_{A\times\epsilon 4} = log(2)$). We next change the parameters for the nuisance state to be $\alpha_0 = 0.5, \alpha_S = log(0.80), \alpha_{\epsilon 4} = log(4), \alpha_A = log(3), \alpha_G = log(2), \alpha_{A\times\epsilon 4} = log(2)$. **Figure 5** presents biases in the estimates and **Supplementary Figures 4-5** show the empirical estimates and the approximations.

Supplementary Figure 7: Frequency of the disease state of interest ($D = 1$) and the nuisance disease ($D = 1^*$) when $\beta_0 = 1.5, \beta_S = log(0.80), \beta_{\epsilon 4} = log(8), \beta_A = log(3)$, $\beta_G = log(1), log(1.5), log(2), log(2.5), log(3), \beta_{G\times\epsilon 4} = log(1), log(2), log(3), log(3)$, $\alpha_0 = 0.5, \alpha_S = log(0.80), \alpha_{\epsilon 4} = log(4), \alpha_A = log(3), \alpha_G = log(2), \alpha_{A\times\epsilon 4} = log(2)$, $\theta_G = 0.10, \theta_A = 0.50, \theta_S = 0.52, \theta_{\epsilon 4} = 0.07$**.** Shown along the x-axis are values of β_G and indicated by color are values of $\beta_{A\times\epsilon 4}$. We note that these frequencies are similar to those in context of Alzheimer's disease.

ROLE OF THE GENETIC VARIANTS SERVING INNATE IMMUNE SYSTEM IN SUSCEPTIBILITY TO ALZHEIMER's DISEASE

We apply the usual logistic analyses with the clinical diagnosis as an outcome variable to a dataset collected as part of the Alzheimer's Disease Genetics Consortium. We next apply the approximations (7)-(10) and (11)-(14) to see how the genetic estimates change when presence of the nuisance disease state is recognized.

We mapped Illumina Human 660K markers onto human chromosomes using NCBI dbSNP database (https://www.ncbi.nlm.nih.gov/projects/SNP/). Chromosomal 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 165 SNPs to reside in genes serving innate immune system.

The dataset consists of 727 controls and 2,797 cases diagnosed with AD.

We are interested to examine a relationship between the pathologic disease state of AD characterized by presence of amyloid deposition and each of the 165 SNPs serving the innate immune system. We include ApoE $\epsilon 4$ status, age, and sex in the model with an interaction between ApoE and age. The genetic variant is modeled as a Bernoulli variable as an indicator of presence or absence of a minor allele. Age is Bernoulli as well that corresponds to a median split in the dataset.

Table 4 presents estimates of effects of the SNPs obtained using the usual logistic regression model with the clinical diagnosis as an outcome variable in a univariable model (3) and with adjustment for SNP + ApoE $\varepsilon 4$ + Age + Sex (7); and the corresponding models (1-2) and (5-6) that recognize presence of the nuisance disease state. In the univariable setting the empirical bias is estimated as the difference between the main effect estimates obtained in model (3) and model (1-2), and the approximation to the bias is estimated as derived in (4b). In the multivariable setting, the empirical bias is the difference between main effect estimates obtained in model (7) and (5-6), and the approximation is as derived in (8b).

First shown in **Table 4** are 16 estimates with p-value<0.05 after the Benjamini-Hochberg multiple testing adjustment in a univariable model (3) and then added are 13 SNPs with p-value <0.05 in a univariable model (1-2). Across all these SNPs, the approximation was accurate relative to the empirical bias.

DISCUSSION

We've examined a situation when multiple disease states share observed symptoms and hence the clinical diagnosis. Both theoretically and in extensive simulation studies we observed that the magnitude of bias can be substantial in situations when frequency of the nuisance disease state within the clinically diagnosed set varies by the key variables. We derived a simple and general approximation to the relationship between the genetic effect estimates that use the clinical diagnosis as an outcome variable and the estimates that recognize presence of the nuisance disease state.

While the effect of misclassification of the disease status has been examined extensively in statistical literature (Carroll et al, 2006), we extend the literature by deriving a simple and general approximation to the bias in a multivariable setting. The approximation provides a simple formula to assess how elastic the estimates of interest are to the values of parameters in the nuisance risk model. The regression coefficients or plausible ranges for the coefficients of the nuisance disease state are often available in the literature.

Simulation studies that we conducted showed that when presence of the nuisance disease is ignored, the genetic effect estimates can be biased in either direction. These biases can be substantial in magnitude leading to false positive and false negative results.

While our study is motivated by the setting of Alzheimer's disease, the results are readily applicable for other complex diseases. For example, Manchia el al (2013) examined the effect of heterogeneity, i.e. presence of non-cases, in the context of diabetes and showed that ignoring the heterogeneity leads to reduced statistical power to detect an association and also reduced the estimated risks attributable to susceptibility alleles.

The approximation that we've derived is widely applicable in other areas of research where the diagnosis is heterogeneous. For example, when disease states correspond to subtypes of a complex disease. We also see the application to the analyses of Electronic Health Records, where the disease status might be subject to exposure-dependent differential misclassification (Chen et al, 2017).

ACKNOWLEDGEMENTS

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)

We thank Ivan Belousov for help with the computations.

LITERATURE CITATIONS

Carroll RJ, Ruppert D, Stefanski, LA, Crainiceanu (2006) Measurement error in nonlinear models: a modern perspective, Second Edition, Chapman and Hall/CRC

Chen Y, Wang J, Chubak J, Hubbard RA (2018) Inflation of type I error rates due to differential misclassification in HER-derived outcomes: empirical illustration using breast cancer recurrence, *Pharmacoepidemiiol Drug Saf,* 2018: 1-5

Kullback S (1959) Information theory and statistics. New York: John Wiley

Lobach I, Sampson J, Alexeyenko A, Lobach S, Zhang L (2018) Case-control studies of gene-environment interactions. When a case might not be the case. *PLOS One,* in press

Manchia M, Cullis J, Gustavo T, Rouleau GY, Uher R, Alda M. (2013) The impact of phenotypic and genetic heterogeneity on results of genome-wide association studies of complex diseases. PLOS One. 2013; 8(10): e76295.

Salloway S and Sperling R (2015) Understanding conflicting neurological findings in patients clinically diagnosed as having Alzheimer Dementia. *JAMA Neurology,* 72 (10): 1106-8

Spinka C, Carroll RJ, Chatterjee N (2005) Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity, *Genetic Epidemiology,* 29(2) 108-127

APPENDIX

A1. Approximation using Kullback-Leibler divergence

We show schematics of the derivations based on Model 3, the other models can be derived accordingly. We denote the model the true model (9)-(10) based on probability $pr_\Gamma(D^{CL}|G,X,Z)$ or $pr_\Gamma(D^{CL},G|X,Z,\delta=1)$ as $Q_{B,A}(D^{CL},G,X,Z)=pr_{B,A}(D^{CL},G|X,Z,\delta=1)$. Similarly, we denote model (3) with (4) as $Q_\Gamma(D^{CL},G,X,Z)=pr_\Gamma(D^{CL}|G,X,Z)$. Kullback (1959) showed that parameters Γ converge to values that minimize Kullback-Leibler divergence criteria between the two models, specifically

$$\gamma = argmin\left\{E_{G,X,Z}\left(E_{D^{CL}|G,X,Z}\left[log\left\{\frac{Q_{B,A}(D^{CL},G,X,Z)}{Q_\Gamma(D^{CL},G,X,Z)}\right\}\right]\right)\right\}.$$

Considerable algebraic derivations arrive to the following system of equations to be solved for parameters Γ

$$E_{G,X,Z}\left[\frac{[pr_{B,A}(D=1|G,X,Z)+pr_{B,A}(D=1^*|G,X,Z)]\times pr(G)}{pr_\Gamma(D^{CL}=1|G,X,Z)}\right.$$
$$\times\frac{\partial}{\partial\Gamma}\frac{exp(\gamma_0+\gamma_X\times X+\gamma_G\times G+\gamma_Z\times Z+\gamma_{X\times Z}\times X\times Z)}{1+exp(\gamma_0+\gamma_X\times X+\gamma_G\times G+\gamma_Z\times Z+\gamma_{X\times Z}\times X\times Z)}$$
$$+\frac{pr_{B,A}(D=0|G,X,Z)\times pr(G)}{pr_\Gamma(D^{CL}=0|G,X,Z)}$$
$$\left.\times\frac{\partial}{\partial\Gamma}\frac{1}{1+exp(\gamma_0+\gamma_X\times X+\gamma_G\times G+\gamma_Z\times Z+\gamma_{X\times Z}\times X\times Z)}\right]=0$$

(A1)

Define $M(X,G,Z;\Gamma)=pr(G)\times\frac{exp(\gamma_0+\gamma_X\times X+\gamma_G\times G+\gamma_Z\times Z+\gamma_{X\times Z}\times X\times Z)}{1+exp(\gamma_0+\gamma_X\times X+\gamma_G\times G+\gamma_Z\times Z+\gamma_{X\times Z}\times X\times Z)}$

Then (A1) becomes

$$E_{G,X,Z}\left[X\times M(X,G,Z;\Gamma)\left\{\frac{pr_{B,A}(D=1|G,X,Z)+pr_{B,A}(D=1^*|G,X,Z)}{pr_\Gamma(D^{CL}=1|G,X,Z)}\right.\right.$$
$$\left.\left.-\frac{pr_{B,A}(D=0|G,X,Z)}{pr_\Gamma(D^{CL}=0|G,X,Z)}\right\}\right]=0$$

$$E_{G,X,Z}\left[Z\times M(X,G,Z;\Gamma)\left\{\frac{pr_{B,A}(D=1|G,X,Z)+pr_{B,A}(D=1^*|G,X,Z)}{pr_\Gamma(D^{CL}=1|G,X,Z)}\right.\right.$$
$$\left.\left.-\frac{pr_{B,A}(D=0|G,X,Z)}{pr_\Gamma(D^{CL}=0|G,X,Z)}\right\}\right]=0$$

$$E_{G,X,Z}\left[G\times M(X,G,Z;\Gamma)\left\{\frac{pr_{B,A}(D=1|G,X,Z)+pr_{B,A}(D=1^*|G,X,Z)}{pr_\Gamma(D^{CL}=1|G,X,Z)}\right.\right.$$
$$\left.\left.-\frac{pr_{B,A}(D=0|G,X,Z)}{pr_\Gamma(D^{CL}=0|G,X,Z)}\right\}\right]=0$$

$$E_{G,X,Z}\left[X\times Z\times M(X,G,Z;\Gamma)\left\{\frac{pr_{B,A}(D=1|G,X,Z)+pr_{B,A}(D=1^*|G,X,Z)}{pr_\Gamma(D^{CL}=1|G,X,Z)}\right.\right.$$
$$\left.\left.-\frac{pr_{B,A}(D=0|G,X,Z)}{pr_\Gamma(D^{CL}=0|G,X,Z)}\right\}\right]=0$$

Values of Γ such that

$$\frac{pr_{B,A}(D=1|G,X,Z)+pr_{B,A}(D=1^{*}|G,X,Z)}{pr_{\Gamma}(D^{CL}=1|G,X,Z)}=\frac{pr_{B,A}(D=0|G,X,Z)}{pr_{\Gamma}(D^{CL}=0|G,X,Z)}=1$$

for all G,X,Z solve the system of equations (A1).

By definition,

$$\gamma_G = 0.25 \times \sum_{x,z}[logit\{pr_\Gamma(D^{CL}=1|G=1,X=x,Z=z)\} - logit\{pr_\Gamma(D^{CL}=1|G=0,X=x,Z=z)\}].$$

With Taylor series expansion around $\beta_G = 0$ we arrive at (12a). Derivation for the other parameters is similar. If X is continuous, then e.g.,

$$\gamma_X = 0.5 \times \sum_{g}[logit\{pr_\Gamma(D^{CL}=1|G=g,X=x+1,Z=0)\} - logit\{pr_\Gamma(D^{CL}=1|G=g,X=x,Z=0)\}].$$

$\alpha_G =$	$Log(1) = 0$	$Log(1.5) = 0.41$	$Log(2) = 0.69$
$n_0 = n_1 = 3{,}000$	0.052	0.72	0.99
$n_0 = n_1 = 10{,}000$	0.048	0.79	0.99

Table 1: False positive rate defined as the proportion of p-values≤0.05 across 10,000 simulated datasets in the usual logistic regression analyses as an outcome variable (3), when in fact $\beta_G = 0$ and the data are generated from (1)-(2). We let $\beta_0 = 0.5$, $\beta_G = 0$, $\alpha_G = \log(1) = 0\,, \log(1.5) = 0.41, \log\,(2) = 0.69$.

α_G	β_G				
	Log(1)=0	Log(1.5)=0.41	Log(2)=0.69	Log(2.5)=0.92	Log(3)=1.1
Log(1)=0	0.003, *0*	0.23, *0.22*	0.41, *0.40*	0.57, *0.56*	0.70, *0.69*
Log(2)=0.41	0.41, *0.41*	0.57, *0.56*	0.70, *0.69*	0.82, *0.81*	0.92, *0.92*
Log(3)=0.69	0.70, *0.69*	0.82, *0.81*	0.93, *0.92*	1.0, *1.0*	1.1, *1.1*
Log(4)=1.1	0.93, *0.92*	1.02, *1.01*	1.1, *1.1*	1.2, *1.2*	1.3, *1.3*

Table 2: Empirical estimates of β_G and *approximation* (4b). The data are simulated from models (1)-(2) and is estimated using model (3). Empirical estimates are the averages across 500 datasets with 3,000 cases and 3,000 controls. We let $\beta_0 = \alpha_0 = 0.5$, $\beta_G = \log(1) = 0\,, \log(1.5) = 0.41\,, \log(2) = 0.69\,, \log(2.5) = 0.92\,, \log(3) = 1.1, \beta_{\epsilon 4} = \alpha_{\epsilon 4} = \log(8)\,, \alpha_G = \log(1) = 0, \log(2) = 0.41\,, \log(3) = 0.69, \log(4) = 1.1.$

α_G	β_G				
	Log(1)=0	Log(1.5)=0.41	Log(2)=0.69	Log(2.5)=0.92	Log(3)=1.1
Log(1)=0	0.0056, *0*	0.23, *0.22*	0.41, *0.40*	0.57, *0.56*	0.70, *0.69*
Log(2)=0.41	0.41, *0.41*	0.57, *0.56*	0.79, *0.69*	0.82, *0.81*	0.92, *0.92*
Log(3)=0.69	0.70, *0.69*	0.82, *0.81*	0.92, *0.92*	1.0, *1.0*	1.1, *1.1*
Log(4)=1.1	0.92, *0.92*	1.02, 1*.01*	1,1, *1.1*	1.2, *1.2*	1.2, *1.3*

Table 3: Empirical estimates of γ_G and *approximation* (8b). The data are simulated from models (5)-(6) and is estimated using model (7). Empirical estimates are the averages across 500 datasets with 3,000 cases and 3,000 controls. We let $\beta_0 = \alpha_0 = 0.5$, $\beta_G = \log(1)\,, \log(1.5)\,, \log(2)\,, \log(2.5)\,, \log(3)\,, \beta_{\epsilon 4} = \alpha_{\epsilon 4} = \log(8)\,, \alpha_G = \log(1)\,, \log(2)\,, \log(3)\,, \log(4)\,.$

Model used for estimation	**SNP only**						**SNP + ApoE ε4 + Age + Sex**					
	(3)		**(1-2)**		**Bias**		**(7)**		**(5-6)**		**Bias**	
SNP	**Estimate**	**P-Value**	**Estimate**	**P-value**	**Empirical**	**Approximation (4b)**	**Estimate**	**P-Value**	**Estimate**	**P-value**	**Empirical**	**Approximation (8b)**
SNPs with p-value <0.05 in the univariable model (3)												
rs906227	1.2	*0.038*	**1.2**	0.09	0.008	0.008	2.2	*0.03*	**1.6**	*0.12*	0.60	0.63
rs7582453	-0.22	*0.028*	**2.4**	*0.02*	-2.7	-2.7	-0.24	*0.03*	**-0.09**	*0.24*	-0.15	-0.17
rs402681	-0.19	*0.047*	**2.3**	*0.008*	-2.4	-2.4	-0.19	*0.07*	**0.06**	*0.38*	-0.25	-0.28
rs4896278	0.22	*0.02*	**-1.6**	*0.01*	1.9	1.9	0.23	*0.03*	**0.61**	*0.21*	-0.38	-0.43
rs4521619	0.22	*0.03*	**0.93**	*0.044*	-0.70	-0.70	0.15	*0.18*	**0.26**	*0.13*	-0.11	-0.17
rs11988857	0.18	*0.049*	**2.5**	*0.01*	-2.3	-2.3	0.20	*0.06*	**0.28**	*0.17*	-0.09	-0.10
rs7046061	-0.21	*0.016*	**-0.22**	*0.006*	0.01	0.01	-0.27	*0.005*	**-0.09**	*0.49*	-0.18	-0.20
rs10745937	-0.19	*0.049*	**2.3**	*0.01*	-2.5	-2.5	-0.14	*0.20*	**-0.06**	*0.53*	-0.09	-0.12
rs4758919	-0.24	*0.007*	**0.19**	*0.02*	-0.43	-0.43	-0.25	*0.01*	**-0.19**	*0.19*	-0.06	-0.08
rs4982421	-0.23	*0.008*	**-0.83**	*0.004*	0.59	0.59	-0.22	*0.03*	**-0.60**	*0.21*	0.39	0.40
rs6573553	0.19	*0.04*	**0.16**	*0.02*	-1.3	-1.3	0.20	*0.045*	**0.30**	*0.35*	-0.10	-0.14

rs2239281	0.24	*0.006*	**1.8**	*0.01*	-1.5	-1.5	0.25	*0.01*	**0.38**	*0.28*	-0.13	-0.15
rs1242558	0.22	*0.01*	**1.8**	*0.02*	-1.5	-1.5	0.20	*0.046*	**0.27**	*0.13*	-0.07	-0.08
rs2469206	-1.04	*0.046*	**-2.1**	*0.02*	1.0	1.0	-1.1	*0.069*	**NA**	*NA*	NA	NA
rs1654558	-0.50	*0.03*	**0.59**	*0.006*	-1.1	-1.1	-0.52	*0.047*	**-0.54**	*0.13*	0.01	0.01
rs6056427	0.27	*0.048*	**-0.28**	*0.04*	0.55	0.55	0.28	*0.06*	**-1.4**	*0.03*	1.7	2.0
SNPs with p-value <0.05 for estimate of β_G in the univariable model (1)-(2)												
rs9380764	0.08	*0.61*	**0.38**	*0.006*	-0.42	-0.42	0.25	0.19	**1.1**	*0.04*	-0.85	-0.87
rs957140	-0.18	*0.07*	**-0.15**	*0.042*	-0.048	-0.048	-0.12	0.29	**NA**	*NA*	NA	NA
rs12900401	-1.7	*0.10*	**-1.7**	*0.036*	-0.01	-0.01	-1.6	0.13	**-1.6**	*0.01*	-0.88	-0.78
rs2469206	-0.36	*0.12*	**-0.35**	*0.042*	-0.01	-0.01	-0.19	0.43	**-1.1**	*0.02*	0.92	0.97
rs165810	-0.21	*0.06*	**-0.18**	*0.01*	-0.03	-0.03	-0.05	0.71	**-0.09**	*0.48*	0.04	0.03
rs330773	0.16	*0.14*	**0.68**	*0.048*	-0.79	-0.79	0.15	0.21	**0.09**	*0.48*	0.06	0.08
rs6781037	0.17	*0.06*	**0.19**	*0.04*	-0.03	-0.03	0.21	0.04	**0.19**	*0.36*	0.02	0.01
rs10051127	0.56	*0.36*	**1.7**	*0.04*	-1	-1	0.45	0.46	**-0.17**	*0.37*	0.62	0.74
rs2402789	0.36	*0.22*	**1.6**	*0.04*	-1.3	-1.3	0.50	0.14	**1.3**	*0.01*	-0.80	-0.90
rs1859333	0.13	*0.14*	**0.23**	*0.03*	-0.03	-0.03	0.12	0.23	**-0.19**	*0.37*	0.31	0.40
rs2283379	0.14	*0.12*	**0.15**	*0.04*	-0.002	-0.002	0.21	0.05	**-0.14**	*0.38*	0.34	0.37
rs17117337	-0.10	*0.46*	**2.2**	*0.004*	0.10	0.10	-0.06	0.67	**-0.15**	*0.38*	0.08	0.08
rs1702447	0.21	*0.16*	**0.58**	*0.006*	-0.08	-0.08	0.22	0.20	**0.20**	*0.24*	0.02	0.03

Table 4: Main effect estimates of SNPs obtained using the usual logistic regression with the clinical diagnosis as an

outcome variable in a univariable model (3) and with adjustment for SNP + ApoE $\varepsilon4$ + Age + Sex (7); and the corresponding models (1-2) and (5-6) that recognize presence of the nuisance disease state. In the univariable setting the empirical bias is estimated as the difference between the main effect estimates obtained in model (3) and model (1-2), and the approximation is as derived in (4b). In the multivariable model, the empirical bias is the difference between main effect estimates obtained in model (7) and (5-6), and the approximation is as derived in (8b).