CASE-CONTROL VS. CASE-ONLY ESTIMATES OF GENE-ENVIRONMENT INTERACTIONS WITH COMMON AND MISCLASSIFIED CLINICAL DIAGOSIS

Iryna Lobach[1], Ying Sheng[1], Siarhei Lobach[2], Lydia Zablotska[1], Chiung-Yu Huang[1]

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

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

*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

ABSTRACT

Genetic studies provide valuable information to assess if the effect of genetic variants varies by the non-genetic ("environmental") variables, what is traditionally defined to be gene-environment interaction. A common complication is that multiple disease states present with the same set of symptoms, and hence share the clinical diagnosis. Because 1) disease states might have distinct genetic bases; and 2) frequencies of the disease states within the clinical diagnosis vary by the environmental variables, analyses of association with the clinical diagnosis as an outcome variable might result in false positive or false negative findings. We develop estimates for assessment of GxE in a case-only study and compare the case-control and case-only estimates. We report extensive simulation studies that evaluate empirical properties of the estimates and show the application to a study of Alzheimer's disease.

INTRODUCTION

Numerous genome-wide association studies (GWAS) have been conducted to estimate how the effect of genetic variants varies by non-genetic (environmental) variables, what is traditionally referred to as gene-environment interactions (GxE). A major and commonly overseen complication is that multiple distinct pathophysiologic mechanisms might present with the same set of symptoms and hence the same clinical diagnosis. Frequencies of the disease states within the clinically diagnosed set often vary by the environmental variables, such as age, race/ethnicity. The pathophysiologic mechanisms underlying the disease states might have distinct genetic bases. Hence the analyses with the clinical diagnosis as an outcome variable might miss important associations or result in spurious findings (Carroll et al, 2016).

Our study is motivated by the setting of Alzheimer's disease (AD) where approximately 30% of patients clinically diagnosed with AD do not have evidence of amyloid deposition as measured by positron emission tomography (PET) (Ossenkoppele et al, 2015). Hence two disease states – symptoms of AD with amyloid evidence and symptoms of AD with no amyloid evidence – are within the clinical diagnosis of AD. We define the disease state of AD with amyloid evidence to be the disease state of interest (Potter and Wisniewski, 2012). Frequencies of the disease states within the clinical diagnosis are estimated to vary by age and Apolipoprotein (ApoE) $\varepsilon 4$ status (Ossenkoppele et al, 2015). Interestingly, ApoE $\varepsilon 4$ status is the most potent genetic factor found thus far. Genetic studies that define the clinical diagnosis as an outcome variable found that risk

of AD increases and the age at onset decreases with the number of ApoE $\varepsilon 4$ alleles (Corder et al, 1993; Farrer et al, 1997), thus suggesting a GxE. Because the disease states of AD with and without the amyloid evidence might have distinct genetic bases and the mechanism of ApoE $\varepsilon 4$ action might be relevant to the amyloid deposition, we are interested to assess the GxE in the relationship to the disease state of AD underlined by the amyloid evidence.

Our previous studies showed both empirically and theoretically that ignoring heterogeneity of AD diagnosis can lead to severely biased estimates of GxE (Lobach et et al, 2018; Lobach et al, 2019).

We are interested to estimate the GxE from the set of clinically diagnosed cases only, assuming that G and E are distributed independently in the population. This interest is supported by the prior statistical literature showing that when the genotype and environment are distributed independently in the population and when the disease is rare, the GxE can be estimated from a case-only study more efficiently than from a case-control study (Piegorsch et al, 1994). This result, however, is not applicable to our setting both because the disease state and the clinical diagnoses are common, and because the clinical diagnosis is not a *surrogate* of the disease state. We, therefore, are interested to derive what types of GxE can be estimated from the set of clinically diagnosed cases and compare variability of case-control vs. case-only estimates.

Our paper proceeds as follows. We first describe the setting and derivations in the Case-control vs. case-only Estimates section. Next, we evaluate the estimates in empirical studies and describe the setting and results in the Simulation Studies section. The application of the methods is then shown on a large-scale study of Alzheimer's disease. We conclude the paper with brief discussion.

Materials and Methods

Case-control vs. case-only estimates

We consider a study consisting of n_1 cases with a clinical diagnosis and n_0 controls. The data are collected using retrospective sampling scheme, i.e. cases are collected from the population of clinically diagnosed cases and controls are collected from the population of the clinically diagnosed controls. Suppose that what measured is a set of genotypes G and environmental variables E that are distributed independently in the population. We define the observed clinical diagnosis be $D^{CL} = \{0,1\}$ and the true disease state to be $D = \{0,1\}$. We let $\pi^{d^{cl}}$ be frequency of the clinical diagnosis in the population, and π^d –the frequency of the disease state in the population. We define frequencies of the disease state of interest within the clinical diagnosis as $pr(D = 1|D^{CL} = 1, G = g, E = e) = \tau(g, e)$. For clarity of the presentation we will assume that the set of controls is homogeneous, i.e. $pr(D = 0|D^{CL} = 0, E = e, G = g) = 1$. For clarity of the presentation we suppose that all variables are binary.

We define $Q(g|\theta)$ be the distribution of genotype in the population according to Hardy-Weinberg Equilibrium.

We define frequencies of the genotype and environment within the clinical diagnosis and the disease states to be $p_{d^{CL}ge} = pr(G = g, E = e|D^{CL} = d^{CL})$ and $\pi_{dge} = pr(G = g, E = e|D = d)$.

We are interested to assess GxE. The traditional analyses are based on the logistic regression model, where GxE is a multiplicative interaction capturing the deviation from the sum of main effects of G and E. We hence start by considering a logistic regression models with the disease state of interest as an outcome variable, where the interaction term is of the primary interest. In the context of this study we are not interested in estimating the main effects and hence the risk model itself, we are just aiming to assess if the data provides sufficient evidence for an interactive effect. Hence consider a model

$$\text{logit}\{pr_B(D = 1|G, E)\} = \beta_0 + \beta_G \times G + \beta_E \times E + \beta_{G\times E} \times G \times E. \quad (1)$$

The disease states, however, are not observed directly, instead what is measured is a clinical diagnosis D^{CL} defined based on the set of observed symptoms. Hence the observed data allows us to estimate an interaction term from the following model

$$\text{logit}\{pr_\Gamma(D^{CL} = 1|G, E)\} = \gamma_0 + \gamma_G \times G + \gamma_E \times E + \gamma_{G\times E} \times G \times E. \quad (2)$$

If the clinical diagnosis is a *surrogate* of the disease state of interest, i.e. $\text{pr}(D^{CL} = d^{cl}|D = d, G = g, E = e) = \text{pr}(D^{CL} = d^{cl}|D = d)$, then $\widehat{\gamma_{G\times E}}$ is a consistent estimate of $\beta_{G\times E}$. In this setting $pr_\Gamma(D^{CL} = d^{cl}|G, E) = \sum_{d^\blacksquare} \text{pr}(D^{CL} = d^{cl}|D = d^\blacksquare) \times pr_B(D = d^\blacksquare|X, E)$, hence if there is no relationship between (G, E) and D, neither there is one between (G, E) and D^{CL}. Otherwise, the probabilities of the clinical diagnosis are

weighted sums of frequencies of the true diagnosis, $\mathrm{pr}(D^{CL} = d^{cl}) = \sum_{d^{\blacksquare}} \mathrm{pr}(D^{CL} = d^{cl}|D = d^{\blacksquare}) \times \mathrm{pr}(D = d^{\blacksquare})$, and $pr_{\Gamma}(D^{CL} = d^{cl}|G, E) = \sum_{d^{\blacksquare}} pr_{\Gamma}(D^{CL} = d^{cl}|D = d^{\blacksquare}, G = g, E = e) \times pr_{B}(D = d^{\blacksquare}|G, E)$. Then if there is no relationship between (G, E) and D, there might be the relationship between (G, E) and D^{CL}.

The seminal work by Piegorsch er al (1994) developed a multiplicative interaction for a rare disease assuming independence between G and E, i.e. for $\beta_{G\times E}$, as follows. An odds ratio (OR) for E in $G = 0$ is then $OR_{10} = \frac{\pi_{000}\times\pi_{101}}{\pi_{001}\times\pi_{100}}$, for G in $E = 0$ is $OR_{01} = \frac{\pi_{000}\times\pi_{110}}{\pi_{010}\times\pi_{100}}$ and for $G = 1, E = 1$ vs. $G = 0, E = 0$ $OR_{11} = \frac{\pi_{000}\times\pi_{111}}{\pi_{011}\times\pi_{100}}$. A multiplicative interaction is then $\Psi = \frac{OR_{11}}{OR_{10}\times OR_{01}}$ and can be estimated in a case-control study as

$$\beta_{CC}^{D} = \log\left(\frac{\pi_{001}\times\pi_{010}\times\pi_{100}\times\pi_{111}}{\pi_{000}\times\pi_{011}\times\pi_{101}\times\pi_{110}}\right). \quad (3)$$

The case-only estimate is

$$\beta_{CO}^{D} = \log\left(\frac{\pi_{100}\times\pi_{111}}{\pi_{101}\times\pi_{110}}\right), \quad (4)$$

which is not applicable to our study for two reasons. First, because the disease states and the clinical diagnosis are not rare. Second, because some of the clinically diagnosed cases are misdiagnosed controls.

Because $pr(E = e, G = g|D = d) \times pr(D = d) = \sum_{d^{CL}} pr(D = d|D^{CL} = d^{CL}, E = e, G = g) \times pr(E = e, G = g|D^{CL} = d^{CL}) \times pr(D^{CL} = d^{CL})$, the GxE interaction in (1), i.e. see (3), becomes

$$\beta^{D}_{G\times E,CC} = \log\left(\frac{\left[p_{001}\times(1-\pi^{dCl})+\{1-\tau(0,1)\}\times p_{101}\times\pi^{dCl}\right]\times\left[p_{010}\times\left(1-\pi^{dCl}\right)+\{1-\tau(1,0)\}\times p_{110}\times\pi^{dCl}\right]}{\left[p_{000}\times(1-\pi^{dCl})+\{1-\tau(0,0)\}\times p_{100}\times\pi^{dCl}\right]\times\left[p_{011}\times(1-\pi^{dCl})+\{1-\tau(1,1)\}\times p_{111}\times\pi^{dCl}\right]}\times \frac{\tau(0,0)\times\tau(1,1)}{\tau(0,1)\times\tau(1,0)}\times\frac{p_{100}\times p_{111}}{p_{101}\times p_{110}}\right). \quad (5)$$

The case-control GxE estimate (5) cannot be seamlessly reduced to a case-only estimate following the arguments of Piegorsh et al (1994) mainly because the disease state and the clinical diagnosis are not rare.

Hence, we are interested to derive other estimates that characterize how the effect of genotype varies by the environment and that can be estimated in a set of clinically diagnosed cases. We aim to derive estimates of GxE that are necessary and might not be sufficient for evaluating whether or not GxE is present.

It can be easily seen that the environment-specific odds of genotype among cases with the disease state of interest is

$$\log\{Odds_G(e)\} = log\left\{\frac{pr(G=1|D=1,E=e)}{pr(G=0|D=1,E=e)}\right\} = log\left\{\frac{\tau(1,e)}{\tau(0,e)}\right\} + log\left\{\frac{pr(G=1|D^{CL}=1,E=e)}{pr(G=0|D^{CL}=1,E=e)}\right\}. \quad (6)$$

Recall that $\theta = pr(G=1)$ and then the environment-specific risk ratio attributable to genotype is

$$\log\{RR_G(E=e)\} = log\left\{\frac{pr(D=1|G=1,E=e)}{pr(D=1|G=0,E=e)}\right\} = log\left\{\frac{\tau(1,e)}{\tau(0,e)}\right\} - \log\left(\frac{\theta}{1-\theta}\right) + log\left\{\frac{pr(G=1|D^{CL}=1,E=e)}{pr(G=0|D^{CL}=1,E=e)}\right\}. \quad (7)$$

From the statistical literature, including the study by Piegorsh et al (1994), we know that a case-only estimate for GxE can be obtained from regressing the environment on the

genotype within the set of cases. i.e. a case-only estimate of GxE is the coefficient α_E from the logistic model

$$log\left\{\frac{pr(G=1|D=1,E=e)}{pr(G=0|D=1,E=e)}\right\} = \alpha_0 + \alpha_E \times e. \quad (8)$$

The coefficient α_E defines GxE on the relative risk scale. Then this model in combination with (6) arrives at

$$log\left\{\frac{pr(G=1|D^{CL}=1,E=e)}{pr(G=0|D^{CL}=1,E=e)}\right\} = \log\left(\frac{\theta}{1-\theta}\right) - log\left\{\frac{\tau(1,e)}{\tau(0,e)}\right\} + \alpha_0 + \alpha_E \times e \quad (9)$$

The analog of model (8) with the clinical diagnosis as an outcome variable is

$$log\left\{\frac{pr(G=1|D^{CL}=1,E=e)}{pr(G=0|D^{CL}=1,E=e)}\right\} = \zeta_0 + \zeta_E \times e. \quad (10)$$

Hence the case-only GxE coefficient α_E can be estimated from the set of clinically diagnosed cases as

$$\alpha_E = \zeta_E + log\left\{\frac{\tau(1,1)}{\tau(0,1)}\right\} - log\left\{\frac{\tau(1,0)}{\tau(0,0)}\right\} \quad (11)$$

We assume that frequency of genotype, θ, and frequencies of the disease states within the clinical diagnosis, $\tau(g,e)$, are known. We define $c_e(g) = \#(G = g|D^{CL} = 1, E = e)$. Then variance of the risk ratios (7) and (11) is

$$Var\left[\log\{\widehat{RR}(E=e)\}\right] = \frac{1}{c_e(1)} + \frac{1}{c_e(0)}. \quad (12)$$

Similarly, variance of the odds (6) is

$$Var\left[\log\{\widehat{Odds}_G(e)\}\right] = \frac{1}{\tau(1,e)\times c_e(1)} + \frac{1}{\tau(0,e)\times c_e(0)}. \quad (13)$$

Remarks:

1. If the clinical diagnosis is rare in the population, i.e. $\pi^{d^{cl}} \approx 0$, then we can see from (5) that the estimate with the clinical diagnosis as an outcome variable is not an unbiased estimate of the GxE for the disease state of interest. That is for rare

diseases ignoring misclassification of the diagnosis does not lead to unbiased estimates of GxE.

2. In Alzheimer's disease study that motivated this work, frequencies of the disease states within the clinical diagnosis are a function of both the genotype and the environment, i.e. $pr(D = 1|D^{CL} = 1, G = g, E = e) = \tau(g, e)$.
3. If the frequency is only a function of the genotype, i.e. $\tau(g, e) = \tau(g)$, or the environment, i.e. $\tau(g, e) = \tau(e)$; then $\alpha_E = \zeta_E$ but the estimate (5) does not reduce to the estimate (3). This setting occurs when the main effect of the environment or genotype is null.
4. We've derived the case-control GxE estimate (5) and proposed three measures to assess how the effect of genotype varies by the environment from a set of clinically diagnosed cases: environment-specific odds of genotype within the subset with the disease state of interest $log\left\{\frac{pr(G=1|D=1,E=e)}{pr(G=0|D=1,E=e)}\right\}$ (6), environment-specific risk ratio attributable to the genotype $\log\{RR_G(E = e)\} = log\left\{\frac{pr(D=1|G=1,E=e)}{pr(D=1|G=0,E=e)}\right\}$ (7), and the coefficient α_E (9).
5. All the risk ratios and odds ratios that we discussed can be easily redefined to be genotype-specific, for example the genotype-specific odds of the environment $\log\{Odds_E(g)\} = log\left\{\frac{pr(E=1|D=1,G=g)}{pr(E=0|D=1,G=g)}\right\} = log\left\{\frac{\tau(g,1)}{\tau(g,0)}\right\} + log\left\{\frac{pr(E=1|D^{CL}=1,G=g)}{pr(E=0|D^{CL}=1,G=g)}\right\}$.

Simulation Studies

We conduct a series of simulation studies to asses bias and false discovery rates of the estimates that ignore presence of the nuisance disease state within the clinical

diagnosis, i.e. estimates (3) and (4) with the clinical diagnosis in place of the disease state. We also assess performance of the odds (6), the risk ratio (7), and the estimate (11); and their variances (12), (13).

In all settings we simulate 500 samples of with n_0 clinically diagnosed controls and n_1 clinically diagnosed cases. We let the genetic variant be Bernoulli(0.10) and the environmental variable be Bernoulli(0.14). We then simulate the disease state D according to the risk model (1) with values of the coefficients $\beta_0, \beta_G, \beta_E, \beta_{G\times E}$ that vary. We next simulate the clinical diagnosis according to frequencies of the nuisance disease state within the clinical diagnosis with $pr(D = 1|D^{CL} = 1, E = e, G = g) = \tau(e, g)$.

We let $\beta_0 = -0.5, \beta_G = \log(1), \log(2), \log(5), \beta_E = \log(2)$. We let $\{pr(D = 1|D^{CL} = 1, E = 0), pr\ (D = 1|D^{CL} = 1, E = 1)\}$ be {0.64, 0.94}; {0.64, 0.84}; {0.84, 0.94}, i.e. we vary both the proportion of the nuisance disease states within the clinically diagnosed set of cases and the difference in the proportion by E.

Setting 1: Null effect, i.e. $\boldsymbol{\beta_{G\times E} = 0}$. We first consider a setting when there is no GxE, i.e. $\beta_{G\times E} = 0$. Shown in **Table 1A** are biases, SDs in the estimates, as well as false discovery rates (FDR). We note that the FDR in the estimates that use the clinical diagnosis as an outcome variable ranges between 0.06 and 0.40 across the settings we considered. Hence these inferences can be substantially inflated. The estimates with the disease state as an outcome variable, i.e. (5) and (11), are nearly unbiased. The case-only estimates tend to have variability that is less or equal to the variability of the

case-control estimates, with relative efficiency varying from 1 to 2.4. Shown in **Table 1B** are the odds (6) and risk ratio (7). The estimates are nearly unbiased with empirical variability that is approximately the same as the theoretical variance (12), as shown in **Table 1C**.

Setting 2: $\boldsymbol{\beta_{G\times E} = log(3)}$. We then consider a setting where $\beta_{G\times E} = \log(3)$. Estimates shown **Table 2A** can be substantially biased when the clinical diagnosis is used in place of the outcome variable, while the bias in nearly removed when the disease state is the outcome. The case-only estimates tend to have variability that is smaller than the variability of the case-control estimates with relative efficiencies varying between 1 and 1.7. Risk ratio estimates in **Table 2B** are nearly unbiased with empirical variability that is close to the theoretical as shown in **Table 2C**.

Alzheimer's disease study

We applied the proposed analyses to a dataset collected as part of the Alzheimer's Disease Genetics Consortium (Naj et al, 2011). 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 $\varepsilon 4$ allele is present in (64.5%) of cases and 365 (29.1%) of controls.

We are considering late-onset AD, hence focus on the subpopulation aged 65 years or older.

Recent biomarker studies (Ossenkoppele et al, 2015) estimate that 95% of ApoE $\varepsilon 4$ carriers and 78% of ApoE $\varepsilon 4$ non-carriers aged 65-75 diagnosed with AD have evidence of amyloid deposition. Similarly, 90% of ApoE $\varepsilon 4$ carriers and 63% of ApoE $\varepsilon 4$ non-carriers aged 75+ clinically diagnosed with AD have evidence of amyloid deposition. We assume that 1 in 10 is diagnosed with AD (https://www.alzheimers.net/resources/alzheimers-statistics/). We will perform sensitivity analyses by assuming these rates and also varying the rates to see how the estimates of ApoE $\varepsilon 4$-by-age interaction might change.

In **Table 3** we present various estimates of how the effect of ApoE $\varepsilon 4$ varies by age with 95% Confidence Intervals (CI) obtained based on 1,000 permutations. We first note that the case-control estimate with the clinical diagnosis as an outcome variable is not statistically significant (-0.08; 95% CI (-0.30, 0.29)), while the case-control estimate with the disease state as an outcome variable is statistically significant (-0.69; 95% CI (-0.76, -0.58)). Across all the settings we considered, length of the 95% CI for the case-only estimates tend to be shorter than the length for the case-control estimate. Setting 1 presented in **Table 3** corresponds to the estimates of the disease states obtained in Ossenkoppele et al, 2015; while Settings 2-4 are slight deviations. We note that the estimates and their 95% CIs are similar across all the settings showing robustness of the conclusions to the slight derivations in the estimates of the disease states within the clinical diagnosis.

Discussion

We've derived case-control and case-only estimates of GxE with respect to the AD disease state underlined by amyloid deposition in the setting when the disease state is not measured, instead frequencies of the disease state within the clinical diagnosis are estimated in a reliability study. We also evaluated bias and false discovery rates for when the misclassification of the clinical diagnosis is ignored. The setting we consider is unique in that the disease states and the clinical diagnosis are common and that rates of the disease state of interest within the clinical diagnosis vary by G and by E.

The development of our study is motivated by the need to assess presence of GxE, as opposed to estimating all parameters in a risk model, e.g. (1). We are interested to estimate the degree to which the effect of genotype varies by the "environment", such as age, sex, education, race/ethnicity, etc. The setting that we've developed offers an advantage of not having to rely on the estimates of main effects.

In simulation experiments we showed that ignoring misclassification of the clinical diagnosis can result in substantial inflation of false positive rates in GxE. Similarly to the original study by Piegorsch et al (1994), we note that the variability of case-only estimates is generally smaller than the variability of the case-control estimates.

The derivations that we've developed rely on estimates of the population frequencies of the disease states of interest within the clinical diagnosis that vary by G and E. These estimates are often available. We advocate for sensitivity analyses by assuming the

estimates obtained in reliability studies and varying the values slightly to see if the GxE estimates change.

While our study is motivated by the setting of Alzheimer's disease, the general development is applicable to other complex diseases and other settings, e.g. analyses of association in the context of the electronic health records, or other types of genetic data, e.g. gene expression. For example, studies of diabetes (Manchia et al, 2013) and analyses of electronic health records (Hubbard et al, 2017).

ACKNOLEDEGMENTS

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)

Literature Citations

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

Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Small, G.W., Roses, A.D., Haines, J.L., and Pericak-Vance, M.A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261, 921–923.

Farrer, L.A., Cupples, L.A., Haines, J.L., Hyman, B., Kukull, W.A., Mayeux, R., Myers, R.H., Pericak-Vance, M.A., Risch, N., and van Duijn, C.M.; APOE and Alzheimer Disease Meta Analysis Consortium (1997). Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. JAMA 278, 1349–1356

Jansen WJ, Ossenkoppele R, Knol DL, Tijms BM, Scheltens P, Verhey FRJ, Visser PJ, and the Amyloid Biomarker Study Group (2015) Prevalence of Cerebral Amyloid Pathology in Persons Without Dementia. A Meta-analysis. JAMA 313(19):1924-1938

Hubbard RA, Johnson E, Chuback J, Wernli KJ, Kamineni A, Bogart A, Ruttr CM (2017) Accounting for misclassification in electronic health records-derived exposures using

generalized linear finite mixture models, *Health Serv Outcomes Research Methods,* 17(2):101-112

Lobach I, Kim I, Alekseyenko A, Lobach S, Zhang L (2019) A simple approximation to bias in the genetic effect estimates when multiple disease states share a clinical diagnosis, *Genetic Epidemiology Journal,* PMID: 30888715

Lobach I, Sampson J, Alekseyenko A, Lobach S, Zhang L (2018) Case-control studies of gene-environment interactions. When a case might not be the case, *PLOS One,* 13(8):e0201140

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,* ONE 8(10): e76295.

Naj, A.C., Jun, G., Beecham, G.W., Wang, L.S., Vardarajan, B.N., et al. Schellenberg, G. (2011) Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with lateonset Alzheimer's. *Nature Genetics*, 43, 436-441

Ossenkoppele R, Jansen WJ, Rabinovici GD, Knol DL, van der Flier WM, van Berckel BNM, Scheltens P, Visser PJ, and the Amyloid PET Study Group (2015) Prevalence of Amyloid PET Positivity in Dementia Syndromes. A Meta-analysis. JAMA 313(19): 1939-1949.

Piegorsch DW, Weinberg CR, Taylor JA (1994) Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies, *Statistics in Medicine,* 13(2): 153-162

Potter H, Wisniewski T. (2012) Apolipoprotein E: essential catalyst of the Alzheimer amyloid cascade. *International Journal of Alzheimer's Disease*. 2012:489428.

$pr(D = 1\|D^{CL} = 1, G = 0)$	$pr(D = 1\|D^{CL} = \|G = 1)$	β_G	Clinical diagnosis as an outcome variable						Disease state is the outcome variable				
			Case-control estimate			Case-only estimate			Case-control estimate		Case-only estimate		
			Bias	SD	False positive rate	Bias	SD	False positive rate	Bias	SD	True value	Bias	SD
0.64	0.94	Log(1)=0	0.21	0.15	0.29	0.20	0.12	0.40	-0.12	0.13	0.00	-0.0002	0.12
		Log(2)=0.69	0.21	0.16	0.30	-0.21	0.12	0.14	-0.11	0.12	-0.12	0.01	0.12
		Log(5)=1.61	0.21	0.18	0.22	-0.03	0.10	0.07	0.09	0.11	-0.25	0.05	0.10
0.64	0.84	Log(1)=0	0.14	0.14	0.15	0.14	0.12	0.22	-0.08	0.12	0.00	-0.0007	0.12
		Log(2)=0.69	0.15	0.16	0.16	0.03	0.11	0.09	-0.02	0.10	-0.13	0.02	0.11
		Log(5)=1.61	0.17	0.27	0.08	-0.10	0.10	0.20	0.12	0.16	-0.24	-0.004	0.10
0.84	0.94	Log(1)=0	0.07	0.15	0.08	0.07	0.11	0.09	-0.03	0.14	0.008	0.003	0.11
		Log(2)=0.69	0.07	0.16	0.09	-0.04	0.11	0.09	-0.002	0.14	0.02	-0.11	0.11
		Log(5)=1.61	0.08	0.20	0.06	-0.17	0.09	0.47	0.07	0.14	-0.24	0.001	0.09

Table 1A: Biases and standard deviations (SD) of case-control and case-only estimates when the clinical diagnosis is used in place of the disease state ((3) and (4)) and when the outcome variable is the disease state ((5) and (11)). We simulated 500 datasets with 3,000 clinically diagnosed cases and 3,000 clinically diagnosed controls where the disease state is simulated according to model (1) with coefficients $\beta_{G\times E} = \log(0)$, $\beta_0 = -0.5, \beta_G = \log(1), \log(2), \log(5), \beta_E = \log(3)$ and the clinical diagnosis is then simulated according to frequencies $pr(D = 1|D^{CL} = 1, G = g, E = e)$.

$pr(D = 1\|D^{CL} = 1, G = 0)$	$pr(D = 1\|D^{CL} = 1, G = 1)$	β_G	Disease state is the outcome variable											
			Log(Odds), $log\left\{\frac{pr(G=1\|D=1,E)}{pr(G=0\|D=1,E)}\right\}$ via (6)						Log(Risk Ratio) $log\left\{\frac{pr(D=1\|G=1,E)}{pr(D=1\|G=0,E)}\right\}$ via (7)					
			$E = 0$			$E = 1$			$E = 0$			$E = 1$		
			True value	Estimate	SD	True value	Estimate	SD	True value	Estimate	SD	True value	Estimate	SD
0.64	0.94	Log(1)=0	-1.8	-1.82	0.09	-1.8	-1.82	0.08	0	-0.001	0.09	0	-0.003	0.08

		Log(2)=0.69	-1.44	-1.45	0.08	-1.56	-1.56	0.07	0.37	0.37	0.08	0.26	0.25	0.07
		Log(5)=1.61	-1.17	-1.12	0.07	-1.36	-1.36	0.07	0.69	0.69	0.07	0.45	0.45	0.07
0.64	0.84	Log(1)=0	-1.81	-1.82	0.09	-1.81	-1.82	0.08	0	-0.0006	0.09	0	-0.0013	0.08
		Log(2)=0.69	-1.44	-1.45	0.08	-1.57	-1.56	0.07	0.37	0.37	0.08	0.26	0.26	0.07
		Log(5)=1.61	-1.13	-1.12	0.07	-1.37	-1.36	0.07	0.45	0.69	0.07	0.69	0.45	0.07
0.84	0.94	Log(1)=0	-1.82	-1.82	0.08	-1.82	-1.82	0.07	-0.008	-0.005	0.08	-0.004	-0.002	0.07
		Log(2)=0.69	-1.44	-1.45	0.07	-1.57	-1.56	0.07	0.37	0.37	0.07	0.26	0.26	0.07
		Log(5)=1.61	-1.13	-1.12	0.07	-1.37	-1.37	0.06	0.69	0.69	0.07	0.45	0.45	0.06

Table 1B: True values, empirical averages and standard deviations (SD) of case-only odds (6) and risk ratio (7) when the outcome variable is the disease state. We simulated 500 datasets with 3,000 clinically diagnosed cases and 3,000 clinically diagnosed controls where the disease state is simulated according to model (1) with coefficients $\beta_{G\times E} = \log(0)$, $\beta_0 = -0.5, \beta_G = \log(1), \log(2), \log(5), \beta_E = \log(3)$ and the clinical diagnosis is then simulated according to frequencies $pr(D = 1|D^{CL} = 1, G = g, E = e)$.

$pr(D=1\|D^{CL}=1,G=0)$	$pr(D=1\|D^{CL}=1,G=1)$	β_G	Disease state is the outcome variable	
			$E=0$	$E=1$
			Theoretical SD	
0.64	0.94	Log(1)=0	0.09	0.08
		Log(2)=0.69	0.08	0.07
		Log(5)=1.61	0.07	0.07
0.64	0.84	Log(1)=0	0.09	0.08
		Log(2)=0.69	0.08	0.07
		Log(5)=1.61	0.07	0.07
0.84	0.94	Log(1)=0	0.08	0.07
		Log(2)=0.69	0.07	0.07
		Log(5)=1.61	0.07	0.06

Table 1C: Empirical averages of the theoretical standard deviations (SD) as described in (12). We simulated 500 datasets with 3,000 clinically diagnosed cases and 3,000 clinically diagnosed controls where the disease state is simulated according to model (1) with coefficients $\beta_{G\times E}=\log(0)$, $\beta_0=-0.5$, $\beta_G=\log(1),\log(2),\log(5)$, $\beta_E=\log(3)$ and the clinical diagnosis is then simulated according to frequencies $pr(D=1|D^{CL}=1,G=g,E=e)$.

			Clinical diagnosis as an	Disease state is the outcome variable

$pr(D = 1\|D^{CL} = 1, G = 0)$	$pr(D = 1\|D^{CL} = 1, G = 1)$	β_G	outcome variable								
			Case-control estimate		Case-only estimate		Case-control estimate		Case-only estimate		
			Bias	SD	Bias	SD	Bias	SD	True value	Bias	SD
0.64	0.94	Log(1)=0	-0.14	0.15	-0.60	0.12	-0.10	0.12	0.36	0.07	0.12
		Log(2)=0.69	0.14	0.17	-0.85	0.10	-0.14	0.13	0.10	0.02	0.10
		Log(5)=1.61	0.18	0.21	-1.08	0.10	0.13	0.12	-0.14	0.00.	0.21
0.64	0.84	Log(1)=0	-0.04	0.16	-0.78	0.11	0.02	0.11	0.36	-0.01	0.11
		Log(2)=0.69	-0.02	0.18	-1.02	0.10	-0.02	0.10	0.09	0.006	0.10
		Log(5)=1.61	-0.95	0.10	-1.22	0.10	-0.008	0.03	0.14	0.03	0.10
0.84	0.94	Log(1)=0	-0.58	0.06	-1.36	0.09	-0.06	0.05	0.17	-0.02	0.09
		Log(2)=0.69	-0.68	0.05	-1.36	0.05	0.04	0.05	0.10	-0.07	0.09
		Log(5)=1.61	-0.58	0.06	-1.36	0.08	0.12	0.05	-0.14	-0.14	0.08

Table 2A: Biases and standard deviations (SD) of case-control and case-only estimates when the clinical diagnosis is used in place of the disease state ((3) and (4)) and when the outcome variable is the disease state ((5) and (11)). We simulated 500 datasets with 3,000 clinically diagnosed cases and 3,000 clinically diagnosed controls where the disease state is simulated according to model (1) with coefficients $\beta_{G\times E} = \log(3)$, $\beta_0 = -0.5$, $\beta_G = \log(1), \log(2), \log(5)$, $\beta_E = \log(3)$ and the clinical diagnosis is then simulated according to frequencies $pr(D = 1|D^{CL} = 1, G = g, E = e)$.

			Disease state is the outcome variable

$pr(D=1\|D^{CL}=1,E=0)$	$pr(D=1\|D^{CL}=1,E=1)$	β_G	Odds $log\left\{\frac{pr(G=1\|D=1,E)}{pr(G=0\|D=1,E)}\right\}$ via (6)						Risk Ratio $log\left\{\frac{pr(D=1\|G=1,E)}{pr(D=1\|G=0,E)}\right\}$ via (7)					
			$E=0$			$E=1$			$E=0$			$E=1$		
			True value	Estimate	SD	True value	Estimate	SD	True value	Estimate	SD	True value	Estimate	SD
0.64	0.94	Log(1)=0	-1.81	-1.82	0.09	-1.45	-1.46	0.07	0.0013	-0.0004	0.09	0.36	0.36	0.07
		Log(2)=0.69	-1.44	-1.44	0.07	-1.34	-1.35	0.07	0.37	0.38	0.07	0.47	0.47	0.07
		Log(5)=1.61	-1.13	-1.12	0.07	-1.13	-1.27	0.07	0.69	0.69	0.07	0.55	0.55	0.07
0.64	0.84	Log(1)=0	-1.82	-1.82	0.09	-1.45	-1.46	0.07	-0.004	-0.002	0.09	0.36	0.36	0.07
		Log(2)=0.69	-1.44	-1.44	0.07	-1.33	-1.35	0.07	0.38	0.38	0.07	0.47	0.47	0.07
		Log(5)=1.61	-1.17	-1.17	0.07	-1.28	-1.28	0.07	0.65	0.65	0.07	0.54	0.54	0.07
0.84	0.94	Log(1)=0	-1.82	-1.98	0.06	-1.45	-1.24	0.06	0.001	0.04	0.06	0.36	0.57	0.06
		Log(2)=0.69	-1.34	-1.24	0.06	-1.44	-1.24	0.06	0.37	0.38	0.06	0.47	0.58	0.06
		Log(5)=1.61	-1.13	-1.13	0.06	-1.27	-1.27	0.06	0.69	0.87	0.06	0.55	0.58	0.06

Table 2B: True values, empirical averages and standard deviations (SD) of case-only odds (6) and risk ratio (7) when the outcome variable is the disease state. We simulated 500 datasets with 3,000 clinically diagnosed cases and 3,000 clinically diagnosed controls where the disease state is simulated according to model (1) with coefficients $\beta_{G\times E}=\log(3)$, $\beta_0=-0.5, \beta_G=\log(1),\log(2),\log(5), \beta_E=\log(3)$ and the clinical diagnosis is then simulated according to frequencies $pr(D=1|D^{CL}=1,G=g,E=e)$.

$pr(D=1\|D^{CL}=1,G=0)$	$pr(D=1\|D^{CL}=1,G=1)$	β_G	Disease state is the outcome variable	
			$E=0$	$E=1$
			Theoretical SD	
0.64	0.94	Log(1)=0	0.09	0.09
		Log(2)=0.69	0.07	0.07
		Log(5)=1.61	0.07	0.07
0.64	0.84	Log(1)=0	0.09	0.07
		Log(2)=0.69	0.07	0.07
		Log(5)=1.61	0.07	0.06
0.84	0.94	Log(1)=0	0.06	0.06
		Log(2)=0.69	0.06	0.06
		Log(5)=1.61	0.07	0.06

Table 2C: Empirical averages of the theoretical standard deviations (SD) as described in (12). We simulated 500 datasets with 3,000 clinically diagnosed cases and 3,000 clinically diagnosed controls where the disease state is simulated according to model (1) with coefficients $\beta_{G\times E}=\log(3)$, $\beta_0=-0.5$, $\beta_G=\log(1),\log(2),\log(5)$, $\beta_E=\log(3)$ and the clinical diagnosis is then simulated according to frequencies $pr(D=1|D^{CL}=1,G=g,E=e)$.

Frequencies of the AD disease state with amyloid evidence within the clinical	Setting 1	Setting 2	Setting 3	Setting 4

diagnosis by age				
$pr(D=1\|D^{CL}=1, Age=65-75, \varepsilon4+)$	95%	95%	90%	100%
$pr(D=1\|D^{CL}=1, Age=75+, \varepsilon4+)$	90%	90%	85%	95%
$pr(D=1\|D^{CL}=1, Age=65-75, \varepsilon4-)$	78%	78%	73%	83%
$pr(D=1\|D^{CL}=1, Age=75+, \varepsilon4-)$	63%	63%	58%	68%
$pr(D^{CL}=1)$	10%	7%	10%	10%
Clinical diagnosis is the outcome variable				
Case-control estimate	-0.08 (-0.30, 0.29)			
Case-only estimate	**-0.69** (-0.76, -0.58)			
Disease state is the outcome variable				
Case-control estimate, (5)	**0.89** (0.83, 0.95)	**0.88** (0.82, 0.94)	**0.91** (0.84, 0.96)	**0.87** (0.81, 0.93)
Case-only estimate, (11)	**0.83** (0.73, 0.93)	**0.82** (0.74, 0.92)	**0.84** (0.75, 0.93)	**0.82** (0.72, 0.90)
Odds $log\left\{\frac{pr(\varepsilon4+\|D=1,Age=65-75)}{pr(\varepsilon4-\|D=1,Age=65-75)}\right\}$, (6)	**0.63** (0.58, 0.69)	**0.63** (0.58, 0.70)	**0.65** (0.59, 0.70)	**0.62** (0.57, 0.68)
Odds $log\left\{\frac{pr(\varepsilon4+\|D=1,Age=75+)}{pr(\varepsilon4-\|D=1,Age=75+)}\right\}$, (6)	**0.12** (0.04, 0.20)	**0.12** (0.04, 0.20)	**0.15** (0.08, 0.23)	**0.10** (0.03, 0.18)
Risk Ratio $log\left\{\frac{pr(D=1\|\varepsilon4+,Age=65-75)}{pr(D=1\|\varepsilon4-,Age=65-75)}\right\}$, (7)	**2.4** (2.4, 2.5)	**2.4** (2.4, 2.5)	**2.5** (2.4, 2.5)	**2.4** (2.3, 2.5)
Risk Ratio $log\left\{\frac{pr(D=1\|\varepsilon4+,Age=75+)}{pr(D=1\|\varepsilon4-,Age=75+)}\right\}$, (7)	**2.3** (1.9, 2.0)	**1.9** (1.9, 2.0)	**2** (1.9, 2.0)	**1.9** (1.8, 2.0)

Table 3: Estimates (95% Confidence Intervals) of how the effect of ApoE $\varepsilon4$ varies by age in the Alzheimer's disease study. Setting 1 is as estimated in the literature. The other Settings are slight deviations from the Setting 1 for analyses of sensitivity.