PLOS ONE

RESEARCH ARTICLE

Case-control studies of gene-environment interactions. When a case might not be the case

Iryna Lobach[1]*, Joshua Sampson[2], Alexander Alekseyenko[3], Siarhei Lobach[4], Li Zhang[1,5,6]

1 Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California, United States of America, **2** National Cancer Institute, National Institutes of Health, Bethesda, MD, United States of America, **3** Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, United States of America, **4** Applied Mathematics and Computer Science Department, Belarusian State University, Minsk, Belarus, **5** Department of Medicine, University of California San Francisco, San Francisco, California, United States of America, **6** Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, California, United States of America

* Iryna.lobach@ucsf.edu

Citation: 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. https://doi.org/10.1371/journal.pone.0201140

Editor: Madepalli K. Lakshmana, Torrey Pines Institute for Molecular Studies, UNITED STATES

Received: March 28, 2018

Accepted: July 8, 2018

Published: August 22, 2018

Data Availability Statement: All the relevant data can be found in the database of genotypes and phenotypes (dbGaP). The study accession number is provided in acknowledgement section (phs000372.v1.p1).

Funding: 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

Abstract

Case-control Genome-Wide Association Studies (GWAS) provide a rich resource for studying the genetic architecture of complex diseases. A key is to elucidate how the genetic effects vary by the environment, what is traditionally defined by Gene-Environment interactions (GxE). The overlooked complication is that multiple, distinct pathophysiologic mechanisms may lead to the same clinical diagnosis and often these mechanisms have distinct genetic bases. In this paper, we first show that using the clinically diagnosed status can lead to severely biased estimates of GxE interactions in situations when the frequency of the pathologic diagnosis of interest, as compared to other diagnoses, depends on the environment. We then propose a pseudo-likelihood solution to correct the bias. Finally, we demonstrate our method in extensive simulations and in a GWAS of Alzheimer's disease.

Introduction

We are interested in using data from a case-control Genome-Wide Association Studies (GWAS) to estimate how an "environmental variable" modifies the effect of a genetic variant on a specific, pathologically defined disease state. However, the complication is that in many GWAS, the cases are a heterogenous group, where multiple distinct pathologically defined disease states have led to a common set of symptoms and a shared clinical diagnosis. In these scenarios, a genetic variant will appear to interact with the environmental variable if the genetic variant affects the pathologically defined disease state of interest and the environmental variable is related to the proportion of cases with that disease state.

The issue of heterogeneity among cases is, perhaps, most pronounced in neurologic and psychiatric disorders, where the clinically defined status is based primarily on descriptive criteria and is typically made in absence of biomarker measurements, imaging data, and biopsies. Our specific motivating study is a GWAS of late-onset Alzheimer's disease (AD), a neurodegenerative disorder

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). The NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by the NIA-funded ADCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P50 AG008702 (PI Scott Small, MD), P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P50 AG005134 (PI Bradley Hyman, MD, PhD), P50 AG016574 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI M. Marsel Mesulam, MD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50 AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P50 AG005131 (PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30 AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P50 AG033514 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), P50 AG047270 (PI Stephen Strittmatter, MD, PhD).

Competing interests: The authors have declared that no competing interests exist.

that is clinically characterized by progressive mental decline. Here, we are interested in identifying genetic variants specifically associated with a high abundance of amyloid deposits and neurofibrially tangles in the brain, which we refer to as "histopathologically defined AD."[1] Specifically, we are interested in whether carrying the ApoE $\varepsilon 4$ variant, which in the study is considered the "environmental variable", modifies the effect of SNPs residing in Toll-Like Receptors (TLR) and Receptor for advanced glycation end products (RAGE) on histopathologically defined AD. Importantly, ApoE $\varepsilon 4$ status is likely to be associated with the proportion of the GWAS cases who have histopathologically defined AD. Recent biomarker studies of AD [2] reported that 36% of ApoE $\varepsilon 4$ non-carriers and 6% in ApoE $\varepsilon 4$ carriers clinically diagnosed with AD do not have evidence of amyloid deposition. We provide a more detailed description of ApoE $\varepsilon 4$, other the risk factors for AD and the heterogeneity of the disease in the Discussion section.

We are interested to test an association between single nucleotide polymorphisms (SNPs) residing in Toll-Like Receptors (TLR) and the true AD diagnosis, i.e. our goal is to identify the genetic that might have lea to amyloid plaques with associated cognitive decline. TLRs play a key role in an innate immune response to invading pathogens and are also important for triggering the adaptive immune responses. Dysregulation of human toll-like receptor function has been shown in aging [3]. Specifically to the etiology of AD, TLRs act through modification of the inflammatory state of microglia/macrophages [1]. Receptor for advanced glycation end products (RAGE) has been identified as receptor for amyloid-beta peptide [4].

There is an extensive literature on how the estimates of the main genetic effect can be biased in situations when disease status is misclassified, i.e. the clinical and pathologic diagnoses do not correspond [5]. We extend the literature by investigating the impact of misdiagnosis on estimates of the Gene-Environment interaction (GxE). In case-control studies, the effects of covariates have been traditionally assessed using logistic regression analysis [6]. Recently, however, Chatterjee and Carroll [7] noticed and proved that the assumptions of Hardy-Weinberg Equilibrium and Gene x Environment independence can be leveraged in the appropriate retrospective analyses to gain statistical efficiency. We adopt the principals derived by Chatterjee and Carroll [7] and develop a pseudo-likelihood model in settings when a case defined based on the clinical diagnosis might not be the case in terms of the *true* diagnosis defined pathophysiologically.

Our paper is organized as follows. First, in the Material and Methods section we present the setting, notation, and proposed pseudo-likelihood approach. Next, the Simulation Experiments section describes the simulation experiments conducted to compare the resulting performance of the proposed method with the performance of standard logistic regression using clinically defined disease. In the same section, we apply our method to the motivating study of AD. The Discussion section concludes the paper.

Materials and methods

We define G be the genotype, e.g. SNPs measured at multiple locations. Let X be the environmental variables that interact with G and let Z be other environmental variables. We assume that the genotype is independent of all environmental variables and the genotypes follows Hardy-Weinberg Equilibrium: $G \sim Q(g,\theta)$. If θ is the frequency of minor allele a when the major allele is A, then the Hardy-Weinberg Equilibrium model [8] according to the number of minor alleles is

$$Pr(G = g|\theta) = \begin{cases} 2 \times \theta \times (1-\theta), & if\ g = Aa \\ \theta^2, & if\ g = aa \\ (1-\theta)^2, & if\ g = AA \end{cases}$$

We define $D^{CL} = \{0, 1\}$ be observed clinical disease status defined based on a set of symptoms. Suppose that the same set of symptoms can be caused by two distinct pathophysiologic mechanisms. Let D be the *true* disease status defined based on the underlying pathology, where $D = 1$ indicates the disease of interest, while $D = 1^*$ is the nuisance disease. For ethical and/or budgetary reasons it might not be possible to measure the underlying pathology, hence D is latent. Instead, an evaluation is performed on a subset of patients or in an external reliability study. We define $\tau(X) = \text{pr}(D = 1|D^{CL} = 1, X)$ to be the frequency of the *true* diagnosis of interest within the clinically diagnosed set that varies by the environment X. We let probabilities of the clinical and *true* diagnoses in the population to be $\pi_{d^{cl}} = \text{pr}(D^{CL} = d^{cl})$ and $\pi_d = \text{pr}(D = d)$, respectively.

The clinical and *true* diagnoses are related $\text{pr}(D^{CL} = d^{cl}) = \sum_{d^{\blacksquare}} \text{pr}(D^{CL} = d^{cl}|D = d^{\blacksquare}) \times \text{pr}(D = d^{\blacksquare})$, which indicates that the probabilities of the clinical diagnosis are weighted sums of frequencies of the *true* diagnoses. If $\text{pr}(D^{CL} = d^{cl}|D = d, X = x, G = g) = \text{pr}(D^{CL} = d^{cl}|D = d)$, then D^{CL} is a *surrogate* of D. In this setting, $\text{pr}(D^{CL} = d^{cl}|G, X) = \sum_{d^{\blacksquare}} \text{pr}(D^{CL} = d^{cl}|D = d^{\blacksquare}) \times \text{pr}(D = d^{\blacksquare}|X, G)$; hence if there is no relationship between (X,G) and D, neither there is one between (X,G) and D^{CL}.

We first consider a binary setting where the risk parameters are defined in terms of $D = 1$ vs. $D = 1^*$ and $D = 0$ combined. Then the risk model is defined in terms of coefficients $\text{B} = (\beta_0, \beta_G, \beta_X, \beta_Z, \beta_{G\times X})$ by

$$\log\left\{\frac{\text{pr}(D = 1|G = g, X = x, Z = z)}{\text{pr}(D = 1^* \text{ or } 0|G = g, X = x, Z = z)}\right\} = \beta_0 + \beta_G \times g + \beta_X \times x + \beta_Z \times z + \beta_{G\times X} \times g \times x. \quad (1)$$

In the second setting that we consider the risk model is defined separately for $D = 1$ vs. $D = 0$ in terms of $\text{B} = (\beta_0, \beta_G, \beta_X, \beta_Z, \beta_{G\times X})$ and for $D = 1^*$ vs. $D = 0$ in terms of $\text{B}^* = (\beta_0^*, \beta_G^*, \beta_X^*, \beta_Z^*, \beta_{G\times X}^*)$ by

$$\log\left\{\frac{\text{pr}(D = 1|G = g, X = x, Z = z)}{\text{pr}(D = 0|G = g, X = x, Z = z)}\right\} = \beta_0 + \beta_G \times g + \beta_X \times x + \beta_Z \times z + \beta_{G\times X} \times g \times x;$$

$$\log\left\{\frac{\text{pr}(D = 1|G = g, X = x, Z = z)}{\text{pr}(D = 1^*|G = g, X = x, Z = z)}\right\} = \beta_0^* + \beta_0^* \times g + \beta_X^* \times x + \beta_Z^* \times z + \beta_{G\times X}^* \times g \times x \quad (2)$$

In Eq (2) B and B^* might share coefficients, e.g. if $\beta_Z = \beta_Z^*$.

The observed data are collected using a case-control design where genetic and environmental variables are measured after the disease status is ascertained. However, the data will be analyzed as a random sample. To facilitate this analysis, we let $\delta = 1$ be an indicator of selection into the study and consider the imaginary Bernoulli sampling with $\text{pr}(\delta = 1|D^{CL} = d^{cl}) \propto n_{d^{cl}}/\pi_{d^{cl}}$. Define $\kappa_{d^{cl}} = \beta_0 + \log(n_{d^{cl}}/\pi_{d^{cl}})$ and $\kappa_{d^{cl}}^* = \beta_0^* + \log\left(\frac{n_{d^{cl}}}{\pi_{d^{cl}}}\right)$ with a parameter set $\Omega = (\kappa_0, \beta_0, \beta_G, \beta_X, \beta_Z, \beta_{G\times X}, \theta)$ For model (1) we define

$$\text{S}(d, d^{cl}, g, x, z; \Omega) = \frac{\exp[I(d = 1) \times \{\kappa_{d^{cl}} + \beta_G \times g + \beta_X \times x + \beta_Z \times z + \beta_{G\times X} \times g \times x\}]}{1 + \exp\{\beta_0 + \beta_G \times g + \beta_X \times x + \beta_Z \times z + \beta_{G\times X} \times g \times x\}} \times Q(g; \theta),$$

and for model (2) we define $\Omega = (\kappa_0, \beta_0, \beta_G, \beta_X, \beta_Z, \beta_{G\times X}, \beta_0^*, \beta_G^*, \beta_X^*, \beta_Z^*, \beta_{G\times X}^*, \theta)$.

$$S(d, d^{cl}, g, x, z; \Omega) = \frac{I(d=1)\times \exp\{\kappa_{d^{cl}} + \beta_G \times g + \beta_X \times x + \beta_Z \times z + \beta_{G\times X} \times g \times x\} + I(d=1^*)\times \exp\{\kappa^*_{d^{cl}} + \beta_0^* \times g + \beta_X^* \times x + \beta_Z^* \times z + \beta_{G\times X}^* \times g \times x\}}{1 + \exp\{\beta_0 + \beta_G \times g + \beta_X \times x + \beta_Z \times z + \beta_{G\times X} \times g \times x\} + \exp\{\beta_0^* + \beta_0^* \times g + \beta_X^* \times x + \beta_Z^* \times z + \beta_{G\times X}^* \times g \times x\}} \times Q(g;\theta).$$

In addition we let $\gamma_{d^{cl}|d}(X) = \mathrm{pr}(D^{CL} = d^{cl}|D = d, X)$.

Consider probability, $\Pr(D^{CL},G|X,Z,\delta = 1)$ and define a function $L(d^{CL},g,x,z;\Omega)$ as follows.

$$L(d^{cl}, g, x, z; \Omega) = \frac{S(0,0,g,x,z;\Omega) + \gamma_{d^{cl}|1}(g) \times S(1, d^{cl}, g, x, z; \Omega)}{\sum_{g^{\blacksquare}, d^{cl\blacksquare}} \{S(0,0,g^{\blacksquare},x,z;\Omega) + \gamma_{d^{cl\blacksquare}|1}(g^{\blacksquare}) \times S(1, d^{cl\blacksquare}, g^{\blacksquare}, x, z; \Omega)\}}. \quad (3)$$

The pseudo-likelihood

$$\prod_{i=1}^{N} L(d_i^{cl}, g_i, x_i, z_i; \Omega) \quad (4)$$

can be used in place of the likelihood function based on arguments provided in the Appendix. Define $\Psi(d^{cl},g,x,z;\Omega)$ to be the derivative of $\log\{L(d^{cl},g,x,z;\Omega)\}$ with respect to Ω and

$$\mathcal{L}_N(\Omega) = \sum_{i=1}^{N} \Psi(D_i^{CL}, G_i, X_i, Z_i; \Omega);$$

$$I = n^{-1} E\left\{\frac{\partial \mathcal{L}_N(\Omega)}{\partial \Omega}\right\};$$

$$\Lambda = \sum_{d^{cl}} \frac{n_{d^{cl}}}{n} E\{\Psi(D_i^{CL}, G_i, X_i, Z_i; \Omega)|D^{CL} = d^{cl}\} \times E\{\Psi(D_i^{CL}, G_i, X_i, Z_i; \Omega)|D^{CL} = d^{cl}\}^T,$$

where all expectations are taken with respect to the actual retrospective sampling scheme. Derivations shown in the Appendix demonstrate that under suitable regularity conditions there is a consistent sequence of solutions to $\mathcal{L}_n(\Omega) = 0$ with the following property

$$n^{\frac{1}{2}}\left(\hat{\Omega} - \Omega\right) \Longrightarrow \mathrm{Normal}\{0, I^{-1}(I - \Lambda)I^{-1}\}.$$

Remark 1: The intercept parameter $\kappa_{d^{cl}}$ is a function of the probability of disease in the population. Hence, if the probability of clinical diagnosis in the population is known or a good bound can be specified, this information can be used while estimating parameters. This cannot be done in the usual logistic regression setting.

Results

Simulation experiments

The goal of the simulation study is to examine potential differences in the effect estimates of the genetic and environmental variables in their relationship to the 1) observed clinical diagnosis using the usual logistic regression model (uLR) and pseudo-likelihood model (pMLE) [7]; and 2) to the *true* disease status by using our pseudo-likelihood approach (pMLE-DX) that takes into account that only a proportion of the clinically diagnosed cases have the true disease. In pMLE-DX parameters are estimated based on Eq (4). Parameters are compared by their Bias and Root Mean Squared Error (RMSE). Simulations are performed using MatLab version R2017a.

In each setting we simulate 500 datasets with $n_0 = n_1 \in \{1000,3000,5000,10000,50000\}$. We let the genotype ($G$) be a Bernoulli random variable with frequency 0.10 to mimic a SNP and

allow its effect to follow a recessive or dominant model. We set our other parameters to be similar to the values observed in our GWAS of AD. The binary variable X = {ε4+,ε4−}, which represents the ApoE ε4 status according to presence or absence of ε4 allele that occurs in approximately 14% of the population.

The proportion of the nuisance disease within the clinical diagnosis is defined as pr(D = 1*|D^{CL} = 1,ε4−) = 0.36 and pr(D = 1*|D^{cl} = 1,ε4+) = 0.06. The clinical diagnosis of late onset AD is defined for ages 65 and older. We simulated age (Z_1) to be Bernoulli with frequency 0.50 e.g. corresponding to a median split. Sex (Z_2) is Bernoulli with frequency 0.52 to reflect what we observed in the motivating data example of AD.

Setting A. We first examine a setting when the nuisance disease and controls are equivalent in that the risk parameters are defined for the disease of interest vs. the combination of controls and nuisance disease as in Eq (1). The risk coefficients are $\beta_0 = -1, \beta_G = 0.406, \beta_{Z_2} = -0.083, \beta_{\varepsilon 4} = 2.079, \beta_{G\times\varepsilon 4} = 0.41$. In this setting, the frequency of the *true* disease status is pr(D = 1) = 46%, pr(D = 1|ε4−) = 40%, pr(D = 1|ε4+) = 82%. Table 1 presents properties of the risk parameter estimates in the datasets with $n_0 = n_1$ = 3,000. Additionally, shown in S1 Table are studies with $n_0 = n_1 \in$ {1000,5000,10000,50000}. When the presence of the nuisance disease is ignored (uLR, pMLE), $\hat{\beta}_{\varepsilon 4}$ and $\hat{\beta}_{G\times\varepsilon 4}$ are biased with elevated RMSE. For example, in a study with $n_0 = n_1$ = 3,000, the bias in $\hat{\beta}_{\varepsilon 4}$ is -0.31 in uLR and pMLE, while the bias is reduced to 0.005 by pMLE-DX. RMSE is 0.33 in uLR and pMLE, while it is reduced to 0.12 by pMLE-DX. Similarly, bias in $\hat{\beta}_{G\times\varepsilon 4}$ is 0.56 in uLR and pMLE, while pMLE-DX reduces the bias by more than half. RMSE of $\hat{\beta}_{G\times\varepsilon 4}$ is 2.5x larger when the presence of the nuisance disease is ignored. Notably, estimates of β_{Z_1} and β_{Z_2} are biased in uLR and pMLE. When sample size

Table 1. Bias and RMSE in parameter estimates when $\beta_{G\times\varepsilon 4} \neq 0$.

Parameters	True value	Clinical disease status is the outcome				With consideration of clinical-pathological diagnoses relationship	
		Usual logistic regression		Pseudo-likelihood method (pMLE)		Pseudo-likelihood method (pMLE-DX)	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
n_0 = 3,000 and n_1 = 3,000							
β_0	-1	0.46	0.46	0.98	0.98	-0.0002	0.07
β_G	0.406	-0.13	0.16	-0.13	0.16	-0.008	0.13
β_{Z_1}	1.098	-0.35	0.35	-0.35	0.35	0.003	0.08
β_{Z_2}	-0.083	0.02	0.06	0.02	0.06	-0.004	0.08
$\beta_{\varepsilon 4}$	2.079	-0.31	0.33	-0.31	0.33	0.005	0.12
$\beta_{G\times\varepsilon 4}$	0.693	0.56	2.4	0.26	0.91	0.22	0.93
Pr(G = 1)		0.10		-0.0004	0.004	0.02	0.02

The Bias and Root Mean Squared Error (RMSE) in parameter estimates from simulations using the usual logistic regression with clinical diagnosis as the outcome (uLR), the pseudo-likelihood approach (pMLE), and our newly proposed pseudo-likelihood approach that accounts for misdiagnosis (pMLE-DX). For these simulations, the study included n_0 = 3000 controls and n_1 = 3000 cases. Frequency of ApoE ε4 allele in the population is 14%. Variables Z_1 and Z_2 are Bernoulli with frequencies 0.50 and 0.52, respectively. Frequency of the *true* disease status is 46% in the population; and is 40% among the subpopulation with no ApoE ε4 alleles, and 82% in the subpopulation with at least one ApoE ε4 alleles. Frequency of nuisance disease within the clinical diagnosis varies by ApoE4 status pr(D = 1*|D^{CL} = 1,ε4−) = 0.36 and pr(D = 1*|D^{CL} = 1,ε4+) = 0.06.

https://doi.org/10.1371/journal.pone.0201140.t001

Table 2. Bias and RMSE in parameter estimates when $\beta_{G\times\varepsilon4} = 0$.

Parameters	True value	Clinical disease status is the outcome				With consideration of clinical-pathological relationship	
		Usual logistic regression		Pseudo-likelihood method		Pseudo-likelihood method	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
$n_0 = 3{,}000$ and $n_1 = 3{,}000$							
β_0	-1	0.45	0.45	0.93	0.93	-0.0004	0.07
β_G	1.099	-0.12	0.15	-0.07	-0.15	0.002	0.13
β_{Z_1}	1.098	-0.33	0.34	-0.33	0.34	0.001	0.08
β_{Z_2}	-0.083	0.02	0.06	0.02	0.06	-0.003	0.08
$\beta_{\varepsilon4}$	2.079	-0.26	0.28	-0.26	0.28	0.007	0.12
$\beta_{G\times\varepsilon4}$	0	0.12	0.41	0.13	0.41	0.04	0.43
Pr(G = 1)	0.10			-0.000	0.004	0.03	0.03

The Bias and Root Mean Squared Error (RMSE) in parameter estimates from simulations using the usual logistic regression with clinical diagnosis as the outcome (uLR), the pseudo-likelihood approach (pMLE), and our newly proposed pseudo-likelihood approach that accounts for misdiagnosis (pMLE-DX). For these simulations, the study included $n_0 = 3000$ controls and $n_1 = 3000$ cases. Frequency of ApoE $\varepsilon4$ allele in the population is 14%. Variables Z_1 and Z_2 are Bernoulli with frequencies 0.50 and 0.52, respectively. Frequency of the *true* disease status is 46% in the population; and is 40% among the subpopulation with no ApoE $\varepsilon4$ alleles, and 82% in the subpopulation with at least one ApoE $\varepsilon4$ alleles. Frequency of nuisance disease within the clinical diagnosis varies by ApoE4 status $\text{pr}(D = 1^*|D^{CL} = 1,\varepsilon4-) = 0.36$ and $\text{pr}(D = 1^*|D^{CL} = 1,\varepsilon4+) = 0.06$.

https://doi.org/10.1371/journal.pone.0201140.t002

increased, the uLR bias in $\hat{\beta}_{G\times\varepsilon4}$ decreased, e.g. the bias is 0.08 in a study with $n_0 = n_1 = 10{,}000$; while the bias in $\hat{\beta}_{\varepsilon4}$ persisted. Across all sample sizes, $\hat{\beta}_G$ is biased by approximately -0.13, whereas considering the nuisance disease nearly eliminated the bias, e.g. to -0.01 in a study with $n_0 = n_1 = 1000$.

We next examine if the presence of the nuisance disease could lead us to erroneously conclude that there was a significant $\hat{\beta}_{G\times\varepsilon4}$ when $\beta_{G\times\varepsilon4} = 0$. Here, we simulated datasets with $\beta_{G\times\varepsilon4} = 0$. Table 2 presents estimates in a study with $n_0 = n_1 = 3000$ and S2 Table is based on studies with $n_0 = n_1 \in \{1000,5000,10000,50000\}$. Estimates of β_0, β_G, β_{Z_1}, $\beta_{\varepsilon4}$, and $\beta_{G\times\varepsilon4}$ are clearly biased when the presence of the nuisance disease is ignored. For example, in a study with $n_0 = n_1 =$ 3,000, pMLE-DX decreased the bias in $\hat{\beta}_{G\times\varepsilon4}$ from 0.12 in uLR and pMLE to 0.04, while RMSE remained approximately the same 0.41 vs. 0.43. Similarly, pMLE-DX reduced the bias in $\hat{\beta}_{\varepsilon4}$ from -0.26 in uLR to 0.007. At the same time, the RMSE of $\hat{\beta}_{\varepsilon4}$ went from 0.28 (uLR, pMLE) to 0.12 (pMLE-DX). Increasing the sample size reduced the uLR bias for $\hat{\beta}_{G\times\varepsilon4}$, e.g. the bias is 0.09 in a study with $n_0 = n_1 = 10{,}000$ but did not alleviate the substantial uLR bias in $\beta_{\varepsilon4}$. Across all sample sizes considered, the uLR estimates of β_G are biased by approximately -0.12, while pMLE-DX reduced the bias to e.g. 0.01 in a study with 1,000 cases and 1,000 controls.

We next consider the effect of underestimating $\text{pr}(D = 1^*|D^{CL} = 1,\varepsilon4+)$ and $\text{pr}(D = 1^*|D^{CL} = 1,\varepsilon4-)$ in the pseudo-likelihood. Here, we simulate data using the parameters specified above, but, when fitting the pseudo-likelihood (S3 Table), set $\text{pr}(D = 1^*|D^{CL} = 1,\varepsilon4-) = 0.3$ and $\text{pr}(D = 1^*|D^{CL} = 1,\varepsilon4+) = 0$, i.e. underestimated by 6%. Naturally, this misspecification introduced bias in some of the estimates and hence increased RMSE. Estimates of $\beta_{\varepsilon4}$ were generally affected more than the estimates of the other parameters. For example, in a study with 3,000 cases and 3,000 controls, bias in $\hat{\beta}_{\varepsilon4}$ increased from 0.005 to -0.66 in pMLE-DX, while RMSE went from 0.12 to 0.67. In estimates of $\beta_{G\times\varepsilon4}$, the bias increased from 0.22 to 0.32, while RMSE went up from 0.93 to 0.94. The bias in $\hat{\beta}_G$ increased to -0.10 in a study with 3,000 cases and 3,000 controls, what has not

reached the level of uLR where the bias is -0.12. Estimates of β_{X_2} remained nearly unbiased with the same RMSE.

We next consider the effect of overestimating pr(D = $1^*|D^{CL}$ = 1,ε4+) and pr(D = $1^*|D^{CL}$ = 1,ε4–) in the pseudo-likelihood (S4 Table). Here, we simulate data using the parameters specified above, but, when fitting the pseudo-likelihood, set pr(D = $1^*|D^{CL}$ = 1,ε4–) = 0.42 and pr (D = $1^*|D^{CL}$ = 1,ε4+) = 0.16, i.e. overestimated by 6%. As expected, this misspecification inflated the bias in the risk estimates. For example, in a study of 3,000 cases and 3,000 controls, bias in $\hat{\beta}_{\varepsilon 4}$ increased from 0.005 to -0.43, while RMSE went from 0.12 to 0.44. Bias in $\hat{\beta}_{G\times\varepsilon 4}$ decreased from 0.22 to 0.17, while RMSE remained the same. Estimates of β_G and β_{X_2} remained nearly unbiased.

Setting B. We next examine a setting when two sets of parameters define the risk of disease, i.e. for $D = 1$ vs. $D = 0$ and $D = 1^*$ vs. $D = 0$ according to the risk model (2). Table 3 ($n_0 = n_1$ = 3,000) and S5 Table present parameter estimates in the setting when $\beta_0 = -1, \beta_0^* = -1.7, \beta_G = -0.69, \beta_G^* = 0, \beta_{Z_1} = 0.10, \beta_{Z_2} = -0.083, \beta_{\varepsilon 4} = 1.3, \beta_{\varepsilon 4}^* = 0.5, \beta_{G\times\varepsilon 4} = 1.099, \beta_{G\times\varepsilon 4}^* = 0, \Pr(G = 1) = 0$. With these parameters, the frequencies of the disease of interest and the nuisance disease are pr(D = 1) = 25.1%, pr($D = 1^*$) = 12.5%, pr(D = 1|ε4+) = 45.4%, pr(D = 1^*|ε4+) = 16.1%, pr(D = 1|ε4–) =20%, pr(D = 1^*|ε4–) = 16.1%. When presence of the nuisance disease is ignored (uLR, pMLE), estimates of β_0,$\beta_{\varepsilon 4}$,$\beta_{G\times\varepsilon 4}$,$\beta_G$ are substantially biased.For example, in a study with 3,000 cases and 3,000 controls, in the bias of uLR for $\hat{\beta}_{\varepsilon 4}$ is -0.22, while pMLE-DX reduced this bias to -0.006; the bias of uLR for $\hat{\beta}_{G\times\varepsilon 4}$ is -0.13, while pMLE-DX reduced this bias to 0.01; the bias of uLR bias for $\hat{\beta}_G$ is 0.30, while pMLE-DX reduced it to 0.005. Biases in uLR persisted for larger sample sizes. If *a priori* evidence is sufficient to set parameters $\beta_{G\times\varepsilon 4}^*$ and β_G^* to 0, when in fact

Table 3. Bias and RMSE in parameter estimates when $\beta_G^* = 0$ and $\beta_{G\times\varepsilon 4}^* = 0$.

Parameters	True value	Clinical disease status is the outcome				With consideration of clinical-pathological diagnoses relationship	
		Usual logistic regression		Pseudo-likelihood method (pMLE)		Pseudo-likelihood method (pMLE-DX)	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
n_0 = 3,000 and n_1 = 3,000							
β_0	-1	0.97	0.97	0.74	0.74	0.02	0.06
β_0^*	-1.7					0.008	0.05
β_G	-0.69	0.30	0.31	-0.39	0.39	0.005	0.10
β_G^*	0					-0.02	0.14
β_{Z_1}	0.10	0.002	0.31	0.004	0.05	0.002	0.05
β_{Z_2}	-0.083	-0.004	0.05	-0.0008	0.05	-0.004	0.05
$\beta_{\varepsilon 4}$	1.3	-0.22	0.24	-0.21	0.23	-0.006	0.10
$\beta_{\varepsilon 4}^*$	0.5					-0.007	0.05
$\beta_{G\times\varepsilon 4}$	0.10	-0.13	0.29	-0.28	0.36	0.01	0.25
$\beta_{G\times\varepsilon 4}^*$	0					0.001	0.11
Pr(G = 1)	0.10			0.05	0.05	0.0001	0.004

The Bias and Root Mean Squared Error (RMSE) in parameter estimates from simulations using the usual logistic regression with clinical diagnosis as the outcome (uLR), the pseudo-likelihood approach (pMLE), and our newly proposed pseudo-likelihood approach that accounts for misdiagnosis (pMLE-DX). For these simulations, the study included n_0 = 3000 controls and n_1 = 3000 cases. Risk of the disease of interest is defined in a set of parameters $\beta_0, \beta_G, \beta_{Z_1}, \beta_{Z_2}, \beta_{G\times\varepsilon 4}$; while the risk of the nuisance disease is parametrized by $\beta_0^*, \beta_G^*, \beta_{\varepsilon 4}^*, \beta_{G\times\varepsilon 4}^*$. Frequency of ApoE ε4 allele in the population is 14%. Variables Z_1 and Z_2 are Bernoulli with frequencies 0.50 and 0.52, respectively. Frequencies of the disease of interest and the nuisance disease are pr(D = 1) = 24.8%, pr(D = 1^*) = 12.5%, pr(D = 1|ε4+) = 43%, pr(D = 1^*|ε4 +) = 16.1%, pr(D = 1|ε4–) = 20%, pr(D = 1^*|ε4+) = 11.6%. Frequency of the nuisance disease within the clinical diagnosis varies by ApoE4 status pr(D = $1^*|D^{CL}$ = 1,ε4–) = 0.36 and pr(D = $1^*|D^{CL}$ = 1,ε4+) = 0.06.

https://doi.org/10.1371/journal.pone.0201140.t003

Table 4. Bias and RMSE in parameter estimates when $\beta^*_G = 0, \beta_{G\times\varepsilon 4} = 0$ and $\beta^*_{G\times\varepsilon 4} = 0$.

Parameters	True value	Clinical disease is the outcome				With consideration of clinical-pathological diagnoses relationship	
		Usual logistic regression		Pseudo-likelihood method (pMLE)		Pseudo-likelihood method (pMLE-DX)	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
n_0 = 3,000 and n_1 = 3,000							
β_0	-1	0.97	0.97	0.75	0.75	0.03	0.06
β^*_0	-1.7					0.01	0.05
β_G	-0.69	0.30	0.31	-0.38	0.39	0.004	0.09
β^*_G	0					-0.01	0.13
β_{Z_1}	0.10	0.002	0.05	0.001	0.09	0.002	0.05
β_{Z_2}	-0.083	-0.004	0.05	-0.003	0.05	-0.004	0.05
$\beta_{\varepsilon 4}$	1.3	-0.22	0.24	-0.22	0.23	-0.006	0.10
$\beta^*_{\varepsilon 4}$	0.5					-0.009	0.06
$\beta_{G\times\varepsilon 4}$	0	-0.23	0.28	-0.23	0.28	0.01	0.25
$\beta^*_{G\times\varepsilon 4}$	0					-0.0008	0.12
Pr(G = 1)	0.10					0.000	0.004

The Bias and Root Mean Squared Error (RMSE) in parameter estimates from simulations using the usual logistic regression with clinical diagnosis as the outcome (uLR), the pseudo-likelihood approach (pMLE), and our newly proposed pseudo-likelihood approach that accounts for misdiagnosis (pMLE-DX). For these simulations, the study included n_0 = 3000 controls and n_1 = 3000 cases. Risk of the disease of interest is defined in a set of parameters $\beta_0, \beta_G, \beta_{Z_1}, \beta_{Z_2}, \beta_{G\times\varepsilon 4}$; while the risk of the nuisance disease is parametrized by $\beta^*_0, \beta^*_G, \beta^*_{\varepsilon 4}, \beta^*_{G\times\varepsilon 4}$. Frequency of ApoE $\varepsilon 4$ allele in the population is 14%. Variables Z_1 and Z_2 are Bernoulli with frequencies 0.50 and 0.52, respectively. Frequencies of the disease of interest and the nuisance disease are pr(D = 1) = 24.8%, pr(D = 1*) = 12.5%, pr(D = 1|ε4+) = 43%, pr(D = 1*|ε4 +) = 16.1%, pr(D = 1|ε4−) = 20%, pr(D = 1*|ε4−) = 11.6%. Frequency of the nuisance disease within the clinical diagnosis varies by ApoE4 status pr(D = 1*|D^{CL} = 1,ε4−) = 0.36 and pr(D = 1*|D^{CL} = 1,ε4+) = 0.06.

https://doi.org/10.1371/journal.pone.0201140.t004

these coefficients are zero, then RMSE of pMLE-DX are further reduced by at least 2-fold (data not shown).

Table 4 and S6 Table present the results in a setting similar to that of Table 3 but when there is no interaction between the genotype and ApoE4 status, i.e. $\beta_{G\times\varepsilon 4} = 0$. Ignoring the nuisance disease in the uLR resulted in bias in the estimate of $\beta_{G\times\varepsilon 4}$ that is -0.23, which might mislead to a conclusion that there is an interactive effect between the genotype and ApoE $\varepsilon 4$ status. The bias persisted for larger sample sizes.

Setting C. We next conducted a simulation study to better understand the underlying nature of the biases in the estimates noted when presence of the nuisance disease is ignored (uLR). For clarity, we simulated all variables to be binary. Variables G,Z_1 and Z_2 are Bernoulli with frequencies 0.10, 0.52 and 0.50, respectively. Risk coefficients are $\beta_0 = -1, \beta_G = \log(1.5) = 0.41, \beta_{Z_1} = 1, \beta_{Z_2} = \log(0.92) = -0.08,\ \beta_{\varepsilon 4} = \log(8) = 2.1, \beta_{G\times\varepsilon 4} = \log(3) = 1.1$. Then we varied values of $\beta_{X_2}, \beta_{\varepsilon 4}$, and $\beta_{G\times\varepsilon 4}$. The relationship between clinical and pathophysiological diagnosis is set to be pr(D = 1*|D^{CL} = 1,ε4−) = 0.36 and pr(D = 1*|D^{CL} = 1,ε4+) = 0.06. We simulated 500 datasets with 3,000 cases and 3,000 controls.

Fig 1 presents a study where $\beta_{\varepsilon 4}$ varies as log(1),log(1.5),log(2),log(2.5),. . .log(8) across the x-axis and β_{Z_2} is color-coded to be 0, 0.5, 1, 1.5. We show in panels A, B, C, D, and E, the biases of $\hat{\beta}_{Z_2}, \hat{\beta}_{Z_1}, \hat{\beta}_{\varepsilon 4}, \hat{\beta}_G$, and $\hat{\beta}_{G\times\varepsilon 4}$, respectively. With increasing value of $\beta_{\varepsilon 4}$, the biases in the main effect estimates of β_{Z_2}, β_{Z_1} and β_G increase. For example, the bias in $\hat{\beta}_G$ reaches -0.10 when $\beta_{\varepsilon 4}$ is log(5). The bias in $\hat{\beta}_{\varepsilon 4}$ and $\hat{\beta}_{G\times\varepsilon 4}$ is even more sensitive to value of $\beta_{\varepsilon 4}$. For example, when

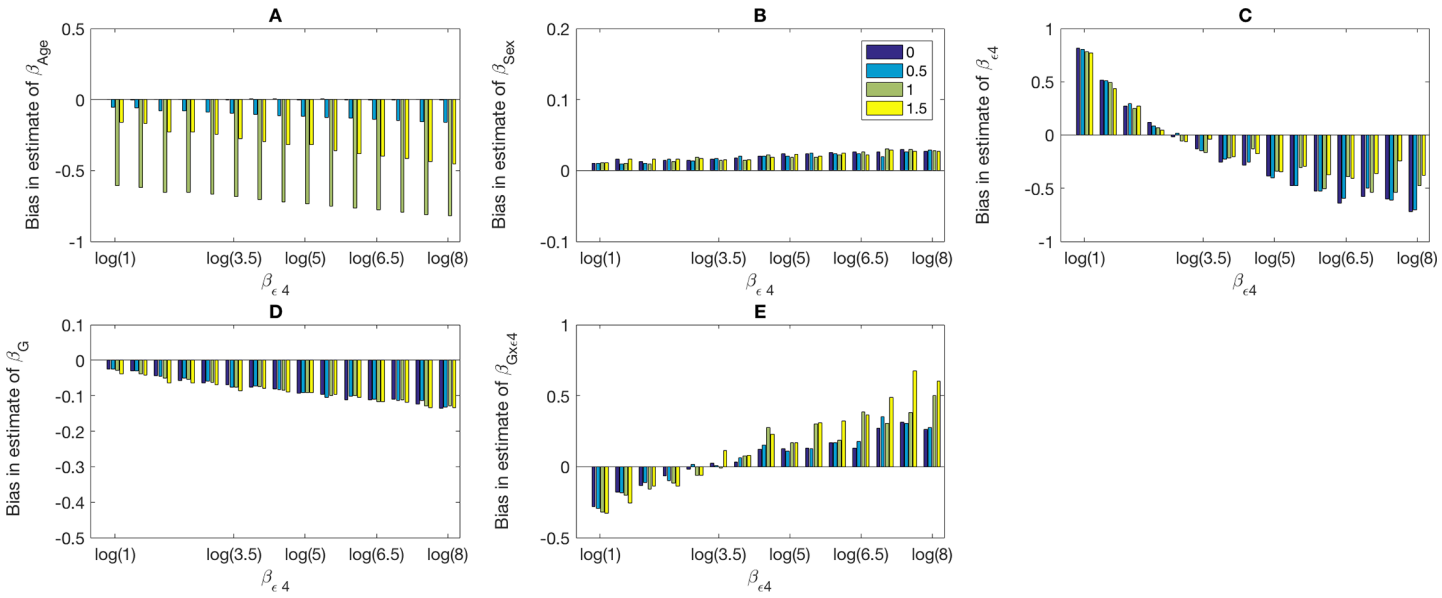

Fig 1. The bias in estimates of β_{Z_1} (β_{Age}) (A), β_{Z_2} (β_{Sex}) (B), $\beta_{\varepsilon 4}$ (C), β_G (D), and $\beta_{G\times\varepsilon 4}$ (E) obtained using the usual logistic regression with clinical diagnosis as the outcome across values of $\beta_{\varepsilon 4}$. Simulated are datasets with 3,000 cases and 3,000 controls. Values of β_{ApoE4} are listed along the x-axis and the true values of β_{Z_1} are indicated by color. The parameters are set as follows: $\beta_0 = -1$, $\beta_G = \log(1.5)$, $\beta_{Z_2} = -0.083$, $\beta_{G\times\varepsilon 4} = \log(3)$; the relationship between the clinical and *true* disease statuses is $\text{pr}(D = 1^* | D^{CL} = 1, \varepsilon 4\text{-}) = 0.36$ and $\text{pr}(D = 1^* | D^{CL} = 1, \varepsilon 4\text{+}) = 0.06$. Variables G,Z_1 and Z_2 are Bernoulli with frequencies 0.10, 0.50 and 0.52, respectively.

https://doi.org/10.1371/journal.pone.0201140.g001

$\beta_{\varepsilon 4} = 0$, the bias in $\hat{\beta}_{G\times\varepsilon 4}$ is 0.8; while when $\beta_{\varepsilon 4} = \log(8)$ the bias is -0.7. Similarly, when $\beta_{\varepsilon 4} = 0$, the bias in $\hat{\beta}_{G\times\varepsilon 4}$ is -0.18; while when $\beta_{\varepsilon 4} = \log(8)$ the bias becomes 0.6. Bias in the estimates of β_{X_2} increases with the increase in the true value. Bias in the other estimates is nearly not affected by values of β_{X_2}.

Fig 2 presents a study where $\beta_{G\times\varepsilon 4}$ varies as log(1),log(1.5),log(2),log(2.5),. . .log(8) across the x-axis and β_{Z_2} is color-coded to be 0, 0.5, 1, 1.5. We show in panels A, B, C, D, and E, the bias of $\hat{\beta}_{Z_2}$, $\hat{\beta}_{Z_1}$, $\hat{\beta}_{\varepsilon 4}$, $\hat{\beta}_G$, $\hat{\beta}_{G\times\varepsilon 4}$, respectively. In this setting, the biases in the main effects $\hat{\beta}_{Z_2}$, $\hat{\beta}_{Z_1}$ and $\hat{\beta}_G$ were approximately the same for all values of $\beta_{G\times\varepsilon 4}$, while the biases in the estimates of $\hat{\beta}_{\varepsilon 4}$ and $\hat{\beta}_{G\times\varepsilon 4}$ were more sensitive to the value of $\beta_{G\times\varepsilon 4}$. For example, when the interaction coefficient is 0, the bias of $\hat{\beta}_{\varepsilon 4}$ is nearly -2, while when $\beta_{G\times\varepsilon 4} = \log(8) = 2.08$, the bias goes up to 3. When $\beta_{G\times\varepsilon 4} = 0$, the bias in the estimate is nearly zero, while the bias goes to almost 6 when the true value is log(8).

Analyses of genetic variants serving toll-like receptors and receptor for advanced glycation end products in Alzheimer's disease

We applied the proposed analyses to a dataset collected as part of the Alzheimer's Disease Genetics Consortium. The data has been anonymized prior to access by the authors. 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.

Illumina Human 660K markers have been mapped onto human chromosomes using NCBI dbSNP database (https://www.ncbi.nlm.nih.gov/projects/SNP/). Chromosome location, proximal gene or genes and gene structure location (e.g. intron, exon, intergenic, UTR) has been recorded for all SNPs. From these data, we inferred 111 SNPs to reside in genes serving Toll-

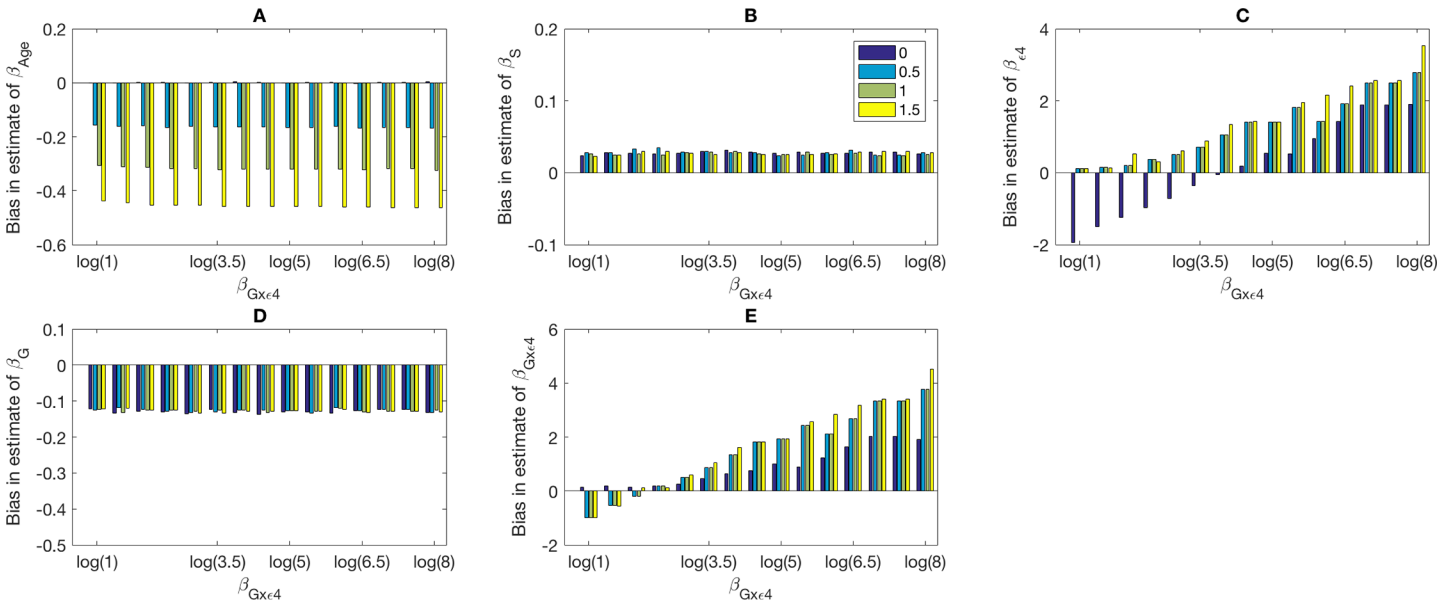

Fig 2. The bias in estimates of β_{Z_1} (β_{Age}) (A), β_{Z_2} (β_{Sex}) (B), $\beta_{\varepsilon 4}$ (C), β_G (D), and $\beta_{G\times\varepsilon 4}$ (E) obtained using the usual logistic regression with clinical diagnosis as the outcome across values of $\beta_{G\times\varepsilon 4}$. Simulated are datasets with 3,000 cases and 3,000 controls. Values of $\beta_{G\times ApoE4}$ are listed along the x-axis and the true values of β_{Z_1} are indicated by color. The parameters are set as follows: $\beta_0 = -1$, $\beta_G = \log(1.5)$, $\beta_{X_2} = -0.083$, $\beta_{G\times\varepsilon 4} = \log(3)$; the relationship between the clinical and *true* disease statuses is $\text{pr}(D = 1^*|D^{CL} = 1,\varepsilon 4\text{-}) = 0.36$ and $\text{pr}(D = 1^*|D^{CL} = 1,\varepsilon 4\text{+}) = 0.06$. Variables G,Z_1 and Z_2 are Bernoulli with frequencies 0.10, 0.50 and 0.52, respectively.

https://doi.org/10.1371/journal.pone.0201140.g002

Like Receptors (TLR). Similarly, we inferred 3 SNPs to reside in the Receptor for advanced glycation end products (AGER).

It is of interest to examine a relationship between the pathologic diagnosis and each of the 111 TLR SNPs (G), ApoE $\varepsilon 4$ status (X), age (Z_1), sex (Z_2). The effect of SNPs might vary by ApoE $\varepsilon 4$ hence we included interaction between the genotype and ApoE $\varepsilon 4$ status. The genetic variables are modeled using a binary indicator of presence or absence of a minor allele.

We estimate parameters using the standard logistic model (uLR) that uses the clinical diagnosis as a surrogate of the pathophysiologic diagnosis and the pseudo-likelihood model (pMLE-DX) where we assume that the relationship between the clinical and pathophysiologic diagnosis is as estimated in the Salloway study [2], i.e. the proportion of the nuisance disease within the clinically diagnosed set is 36% in ApoE $\varepsilon 4$ non-carriers and 6% in ApoE $\varepsilon 4$ carriers. The pseudo-likelihood model pMLE-DX estimates the coefficients in a model that treats the nuisance disease and controls equivalently as in Eq (1). pMLE-DX*, however, estimates two sets of the risk coefficients as in Eq (2). Data analyses are performed using MatLab version R2017a. When optimizing the pseudolikelihood function we bounded the estimates to be on the interval [–5,5].

We first examine the results when statistical significance is assessed according to p-value$<$0.05. We next correct for false discovery rate using Benjanimi-Hochberg method [9].

TLR. Shown in Table 5 are estimates of the risk coefficients for 53 SNPs with permutation-based p-values for $\hat{\beta}_G$ or $\hat{\beta}_{G\times\varepsilon 4}$ that are $<$0.05 in either of the analyses. Of these 53 SNPs, 28 SNPs are within 500k up- or downstream of the SNPs previously reported in GWAS on Alzheimer' disease, dementia, tauopathy, or/and vascular disease (S6 Table).

Estimates of β_G or $\beta_{G\times\varepsilon 4}$ differ numerically between the three approaches. For 14 of these 53 SNPs, $\hat{\beta}_{G\times\varepsilon 4}$ have p-values $<$0.05 in uLR, while in pMLE-DX and pMLE-DX* the corresponding p-values are $>$0.05. These associations detected by uLR might be spurious as a result of clinical-pathophysiological diagnoses relationship varying by ApoE $\varepsilon 4$ status.

Table 5. Parameter estimates in Alzheimer's disease study.

SNP	Gene/Intergenic Region	Method	$\hat{\beta}_{Age}$	$\hat{\beta}_{Sex}$	$\hat{\beta}_{\varepsilon 4}$	$\hat{\beta}_{G}$	$\hat{\beta}_{G\times\varepsilon 4}$
rs2033831	KIAA0922 | TLR2	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**0.86**, p = 0.00	-0.30, p = 0.09	**0.58**, p = 0.03
		pMLE-DX	-1.6, p = 0.13	-2.4, p = 0.11	-2.6, p = 0.32	-0.87, p = 0.33	3.7, p = 0.80
		pMLE-DX*	**-0.24**, p = 0.00	**-0.43**, p = 0.00	1.9, p = 0.25	4.2, p = 0.28	0.22, p = 0.28
rs7656500	KIAA0922 | TLR2	uLR	**-0.25**, p = 0.00	**-0.42**, p = 0.00	**1.6**, p = 0.01	**0.79**, p = 0.01	-0.16, p = 0.43
		pMLE-DX	-1.2, p = 0.25	-1.8, p = 0.27	0.12, p = 0.57	**6.0**, p = 0.02	0.73, p = 0.27
		pMLE-DX*	-0.25, p = 0.00	-0.43, p = 0.00	-2.3, p = 0.27	5.0, p = 0.13	4.3, p = 0.09
rs1816702	TLR2	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.4**, p = 0.00	0.43, p = 0.06	0.01, p = 0.49
		pMLE-DX	-1.4, p = 0.13	-2.1, p = 0.10	-0.18, p = 0.50	**5.0**, p = 0.04	1.1, p = 0.65
		pMLE-DX*	-0.24, p = 0.00	-0.43, p = 0.00	**2.6**, p = 0.03	5.0, p = 0.38	-0.46, p = 0.13
rs830832	SORBS2 | TLR3	uLR	**-0.24**, p = 0.00	**-0.42**, p = 0.00	**0.74**, p = 0.01	-0.31, p = 0.06	**0.74**, p = 0.01
		pMLE-DX	-1.0, p = 0.16	-1.6, p = 0.10	-0.64, p = 0.42	-0.55, p = 0.35	1.8, p = 0.70
		pMLE-DX*	**-0.24**, p = 0.00	**-0.43**, p = 0.00	-1.4, p = 0.07	4.4, p = 0.21	**2.6**, p = 0.03
rs7676342	SORBS2 | TLR3	uLR	**-0.25**, p = 0.00	**-0.43**, p = 0.00	**1.6**, p = 0.00	**0.35**, p = 0.04	-0.24, p = 0.21
		pMLE-DX	-1.3, p = 0.19	-2.2, p = 0.15	0.94, p = 0.72	0.88, p = 0.75	-0.02, p = 0.53
		pMLE-DX*	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.5**, p = 0.03	4.3, p = 0.66	0.23, p = 0.69
rs4862611	SORBS2 | TLR3	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.4**, p = 0.00	0.08, p = 0.24	0.03, p = 0.55
		pMLE-DX	**-1.5**, p = 0.13	**-2.9**, p = 0.09	0.04, p = 0.56	-0.36, p = 0.35	1.2, p = 0.32
		pMLE-DX*	**-0.25**, p = 0.00	**-0.43**, p = 0.00	1.7, p = 0.08	2.2, p = 0.15	**-2.8**, p = 0.03
rs13113778	SORBS2 | TLR3	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**2.0**, p = 0.00	0.13, p = 0.62	-0.64, p = 0.14
		pMLE-DX	-1.7, p = 0.16	-4.9, p = 0.05	4.1, p = 0.10	-2.1, p = 0.26	-3.1, p = 0.25
		pMLE-DX*	**-0.24**, p = 0.00	**-0.42**, p = 0.00	-0.75, p = 0.49	5.0,p = 0.36	2.7, p = 0.06
rs1869617	SORBS2 | TLR3	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**0.96**, p = 0.02	-0.33, p = 0.15	0.46, p = 0.86
		pMLE-DX	-1.4, p = 0.14	-2.7, p = 0.09	-3.1, p = 0.32	-0.53, p = 0.39	4.0, p = 0.11
		pMLE-DX*	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**2.2**, p = 0.00	**4.6**, p = 0.01	0.18, p = 0.51
rs11938703	SORBS2 | TLR3	uLR	**-0.24**, p = 0.002	**-0.42**, p = 0.00	**0.95**, p = 0.00	-0.24, p = 0.05	**0.58**, p = 0.00
		pMLE-DX	-0.78, p = 0.15	-1.2, p = 0.13	0.09, p = 0.57	-0.50, p = 0.28	1.2, p = 0.67
		pMLE-DX*	**-0.25**, p = 0.00	**-0.42**, p = 0.00	1.7, p = 0.20	2.8, p = 0.28	-0.57, p = 0.35
rs1519318	SORBS2 | TLR3	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.2**, p = 0.00	-0.01,p = 0.51	0.21, p = 0.16
		pMLE-DX	-1.4, p = 0.14	-2.6, p = 0.10	0.86, p = 0.70	0.04, p = 0.59	0.07, p = 0.44
		pMLE-DX*	**-0.25**, p = 0.00	**-0.43**, p = 0.00	2.1, p = 0.08	**3.3**, p = 0.01	-0.41, p = 0.39
rs12648771	SORBS2 | TLR3	uLR	**-0.25**, p = 0.00	**-0.42**, p = 0.00	**3.0**, p = 0.00	-0.27, p = 0.26	**-1.6**, p = 0.004
		pMLE-DX	-1.6, p = 0.15	-2.0, p = 0.13	2.9, p = 0.15	-1.6, p = 0.26	-2.1, p = 0.38
		pMLE-DX*	-0.24, p = 0.00	-0.43, p = 0.00	0.62, p = 0.06	5.0, p = 0.24	0.02, p = 0.31
rs6894	NQO1 | LOC100132364	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	0.59, p = 0.07	**-0.64**, p = 0.03	**0.84**, p = 0.03
		pMLE-DX	-1.4, p = 0.12	-2.4, p = 0.09	0.42, p = 0.64	-0.20, p = 0.46	0.49, p = 0.59
		pMLE-DX*	-0.24, p = 0.00	-0.42, p = 0.00	**2.8**, p = 0.00	**3.9**, p = 0.00	0.84, p = 0.16
rs3775296	TLR3	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**2.0**, p = 0.00	-0.02, p = 0.50	-0.61, p = 0.06
		pMLE-DX	-1.4, p = 0.14	-2.4, p = 0.11	3.4, p = 0.14	0.81, p = 0.27	-2.5, p = 0.33
		pMLE-DX*	**-0.24**, p = 0.00	**-0.42**, p = 0.00	**1.4**, p = 0.02	**4.3**, p = 0.046	0.89, p = 0.12
rs7668666	TLR3	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.4**, p = 0.00	-0.14, p = 0.25	0.01, p = 0.52
		pMLE-DX	-1.7, p = 0.13	**-4.9**, p = 0.018	3.4, p = 0.14	1.7, p = 0.19	-2.6, p = 0.27
		pMLE-DX*	**-0.24**, p = 0.00	**-0.42**, p = 0.00	**1.7**, p = 0.02	**4.9**, p = 0.00	-0.81, p = 0.19
rs1706143	TLR3 | FAM149A	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.3**, p = 0.00	-0.12, p = 0.20	0.12, p = 0.26
		pMLE-DX	-1.4, p = 0.14	-2.4, p = 0.11	0.33, p = 0.37	-0.39, p = 0.32	0.66, p = 0.33
		pMLE-DX*	**-0.24**, p = 0.00	**-0.42**, p = 0.00	0.45, p = 0.67	3.8, p = 0.26	**2.9**, p = 0.03
rs9299251	ASTN2 | TLR4	uLR	**-0.25**, 0.00	**-0.43**, p = 0.00	**1.1**, p = 0.00	-0.04, p = 0.65	**0.35**, p = 0.04
		pMLE-DX	-1.3, p = 0.14	-2.1, p = 0.14	0.00, p = 0.63	0.19, p = 0.65	1.1, p = 0.32
		pMLE-DX*	-0.24, p = 0.00	-0.42, p = 0.00	0.48, p = 0.18	3.3, p = 0.45	0.47, p = 0.18
rs955302	TNFRSF19	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.2**, p = 0.00	-0.28, p = 0.06	0.27, p = 0.13
		pMLE-DX	-1.5, p = 0.13	-2.5, p = 0.10	1.2, p = 0.34	-0.23, p = 0.40	-0.37, p = 0.44
		pMLE-DX*	**-0.25**, p = 0.00	**-0.43**, p = 0.00	**2.2**, p = 0.02	3.9, p = 0.30	-1.2, p = 0.16
rs17419570	ASTN2 | TLR4	uLR	**-0.24**, p = 0.00	**-0.44**, p = 0.00	0.68, p = 0.06	**-0.94**, p = 0.01	0.74, p = 0.06
		pMLE-DX	-0.91, p = 0.23	-1.5, p = 0.21	-0.93, p = 0.50	-1.8, p = 0.32	1.9, p = 0.2
		pMLE-DX*	**-0.24**, p = 0.00	**-0.42**, p = 0.00	1.2, p = 0.44	**4.5**, p = 0.008	1.6, p = 0.064

(*Continued*)

Table 5. (Continued)

SNP	Gene/Intergenic Region	Method	$\hat{\beta}_{Age}$	$\hat{\beta}_{Sex}$	$\hat{\beta}_{\varepsilon 4}$	$\hat{\beta}_{G}$	$\hat{\beta}_{G\times\varepsilon 4}$
rs16905625	ASTN2 | TLR4	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.3**, p = 0.01	0.04, p = 0.59	0.16, p = 0.69
		pMLE-DX	-1.5, p = 0.19	-3.3, p = 0.13	3.3, p = 0.12	2.3, p = 0.15	-2.5, p = 0.28
		pMLE-DX*	**-0.25**, p = 0.00	**-0.43**, p = 0.00	0.92, p = 0.11	**3.2**, p = 0.03	-0.33,p = 0.32
rs10513307	ASTN2 | TLR4	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.7**, p = 0.01	0.06, p = 0.56	-0.28, p = 0.31
		pMLE-DX	-1.7, p = 0.15	**-5.0**, p = 0.03	3.9, p = 0.11	-0.61, p = 0.41	-3.0, p = 0.32
		pMLE-DX*	**-0.24**, p = 0.00	**-0.43**,p = 0.00	-1.3, p = 0.33	**5.0**, p = 0.02	2.5, p = 0.09
rs1890047	ASTN2 | TLR4	uLR	**-0.25**, p = 0.002	**-0.43**, p = 0.00	**1.1**, p = 0.00	-0.06, p = 0.34	**0.39**,p = 0.02
		pMLE-DX	-1.3, p = 0.11	-2.2, p = 0.10	-0.77, p = 0.45	-0.08, p = 0.44	1.9, p = 0.22
		pMLE-DX*	**-0.25**, p = 0.00	**-0.43**, p = 0.00	-0.17, p = 0.44	**4.6**, p = 0.008	2.0, p = 0.064
rs4837254	ASTN2 | TLR4	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.2**, p = 0.00	-0.03, p = 0.43	0.27, p = 0.08
		pMLE-DX	-1.7, p = 0.15	**-5.0**, p = 0.03	3.9, p = 0.11	-0.61, p = 0.41	-3.0, p = 0.32
		pMLE-DX*	**-0.24**, p = 0.00	**-0.42**, p = 0.00	**2.0**, p = 0.04	3.4, p = 0.65	-0.90, p = 0.25
rs13285674	ASTN2 | TLR4	uLR	**-0.24**, p = 0.002	**-0.43**, p = 0.00	**1.0**, p = 0.00	**-0.47**, p = 0.003	0.42, p = 0.10
		pMLE-DX	-1.6, p = 0.14	-3.1, p = 0.08	0.31, p = 0.62	-1.3, p = 0.29	0.63, p = 0.40
		pMLE-DX*	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**2.8**, p = 0.00	**4.5**, p = 0.03	-0.65, p = 0.24
rs1337208	ASTN2 | TLR4	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.1**, p = 0.00	-0.009, p = 0.49	**0.34**, p = 0.05
		pMLE-DX	-1.2, p = 0.12	-2.0, p = 0.10	-0.03, p = 0.55	0.15, p = 0.62	1.1, p = 0.31
		pMLE-DX*	**-0.25**, p = 0.00	**-0.42**, p = 0.00	0.96, p = 0.79	2.9, p = 0.15	0.46, p = 0.71
rs1415378	ASTN2 | TLR4	uLR	**-0.25**, p = 0.00	**-0.43**, p = 0.00	**1.4**, p = 0.00	0.05, p = 0.32	0.05, p = 0.54
		pMLE-DX	-1.5, p = 0.14	-2.8, p = 0.10	1.5, p = 0.20	0.98, p = 0.21	-0.62, p = 0.37
		pMLE-DX*	**-0.24**, p = 0.00	**-0.42**, p = 0.00	0.50, p = 0.28	2.1, p = 0.06	1.4, p = 0.08
Rs504204	ASTN2 | TLR4	uLR	**-0.24**, p = 0.002	**-0.43**, p = 0.00	0.21, p = 0.63	0.11, p = 0.54	1.2, p = 0.17
		pMLE-DX	-1.5, p = 0.16	-2.9, p = 0.13	-4.0, p = 0.27	-2.5, p = 0.48	**5.0**, p = 0.044
		pMLE-DX*	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**-3.9** p = 0.16	5.0, p = 0.11	4.6, p = 0.06
rs12337381	ASTN2 | TLR4	uLR	**-0.25**, p = 0.00	**-0.43**, p = 0.00	**0.84**, p = 0.04	0.14, p = 0.66	0.59, p = 0.11
		pMLE-DX	-1.5, p = 0.14	-2.5, p = 0.10	-0.63, p = 0.50	-0.60, p = 0.40	1.6, p = 0.30
		pMLE-DX*	**-0.25,** p = 0.00	**-0.43,** p = 0.00	**3.0**, p = 0.01	4.6, p = 0.08	-1.5, p = 0.09
rs1952464	ASTN2 | TLR4	uLR	**-0.25**, p = 0.00	**-0.43**, p = 0.00	**0.98**, p = 0.00	-0.07, p = 0.34	**0.50**, p = 0.01
		pMLE-DX	-1.5, p = 0.15	-2.3, p = 0.10	-0.08, p = 0.56	0.62, p = 0.33	1.1, p = 0.67
		pMLE-DX*	**-0.25**, p = 0.00	**-0.43**, p = 0.00	**2.9**, p = 0.002	4.6, p = 0.08	-1.5, p = 0.09
rs12342331	ASTN2 | TLR4	uLR	**-0.25**, p = 0.00	**-0.43**, p = 0.00	0.75, p = 0.08	0.07, p = 0.59	0.68, p = 0.08
		pMLE-DX	-1.5, p = 0.13	-2.5, p = 0.11	-0.80, p = 0.50	-0.59, p = 0.38	1.8, p = 0.29
		pMLE-DX*	**-0.25**, p = 0.00	**-0.43**, p = 0.00	**1.4**, p = 0.01	**5.0**, p = 0.03	-0.35, p = 0.09
rs16905754	ASTN2 | TLR4	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.1**, p = 0.01	0.52, p = 0.33	**-1.1**, p = 0.002
		pMLE-DX	-1.5, p = 0.05	-2.7, p = 0.06	5.0, p = 0.07	-0.99, p = 0.41	-4.1, p = 0.21
		pMLE-DX*	**-0.24**, p = 0.00	**-0.43**, p = 0.00	0.11?, p = 0.00	5.0, p = 0.66	0.19, p = 0.08
rs2771054	ASTN2 | TLR4	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**2.9**, p = 0.01	**1.5**, p = 0.01	**-1.5**, p = 0.016
		pMLE-DX	-1.5, p = 0.22	-2.4, p = 0.20	**5.0**, p = 0.04	**4.9**, p = 0.04	-4.1, p = 0.19
		pMLE-DX*	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**-0.19**, p = 0.10	**5.0**, p = 0.15	1.7, p = 0.66
rs521581	ASTN2 | TLR4	uLR	**-0.25**, p = 0.002	**-0.43**, p = 0.00	**1.1**, p = 0.00	-0.02, p = 0.43	**0.34**, p = 0.04
		pMLE-DX	-1.7, p = 0.11	-3.0, p = 0.08	0.35, p = 0.58	0.82,p = 0.33	0.61, p = 0.38
		pMLE-DX*	**-0.25**, p = 0.00	**-0.43**, p = 0.00	1.5, p = 0.10	3.5, p = 0.62	-0.36, p = 0.32
rs1329063	ASTN2 | TLR4	uLR	**-0.24**, p = 0.002	**-0.43**, p = 0.00	**1.6**, p = 0.00	**0.65**, p = 0.02	-0.23, p = 0.33
		pMLE-DX	-1.5, p = 0.19	-2.8, p = 0.15	0.92, p = 0.34	-0.10, p = 0.54	-0.01, p = 0.57
		pMLE-DX*	**-0.24**, p = 0.00	**-0.43**, p = 0.00	-2.0, p = 0.17	5.0, p = 0.58	3.3, p = 0.09
rs495083	ASTN2 | TLR4	uLR	**-0.25**, p = 0.00	**-0.42**, p = 0.00	**1.2**, p = 0.01	-0.19, p = 0.10	0.24, p = 0.10
		pMLE-DX	-1.5, p = 0.11	-3.3, p = 0.08	1.5, p = 0.25	0.37, p = 0.28	-0.72, p = 0.33
		pMLE-DX*	**-0.24**, p = 0.00	**-0.42**, p = 0.00	**4.7**, p = 0.00	4.6, p = 0.10	**-2.8**, p = 0.04
rs476	ASTN2 | TLR4	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**3.0**, p = 0.00	**4.9**, p = 0.03	-1.3, p = 0.11
		pMLE-DX	-1.5, p = 0.13	-3.0, p = 0.07	2.3, p = 0.20	1.1, p = 0.19	-1.6, p = 0.29
		pMLE-DX*	**-0.24**, p = 0.00	**-0.43**, p = 0.00	0.85, p = 0.12	3.2, p = 0.36	0.50, p = 0.64
rs565055	ASTN2 | TLR4	uLR	**-0.25**, p = 0.00	**-0.43**, p = 0.00	**1.1**, p = 0.00	-0.01, p = 0.47	**0.37**, p = 0.01
		pMLE-DX	-1.3, p = 0.16	-2.1, p = 0.11	0.81, p = 0.67	0.15, p = 0.65	0.14,p = 0.57
		pMLE-DX*	**-0.25**, p = 0.002	**-0.42**, p = 0.00	1.0, p = 0.21	0.84, p = 0.59	2.1, p = 0.33

(*Continued*)

Table 5. (Continued)

SNP	Gene/Intergenic Region	Method	$\hat{\beta}_{Age}$	$\hat{\beta}_{Sex}$	$\hat{\beta}_{\varepsilon 4}$	$\hat{\beta}_{G}$	$\hat{\beta}_{G\times\varepsilon 4}$
rs2094630	ASTN2 | TLR4	uLR	**-0.25**, p = 0.00	**-0.43**, p = 0.00	**1.1**, p = 0.00	-0.2, p = 0.00	**0.38**, p = 0.02
		pMLE-DX	-1.3, p = 0.15	-2.1, p = 0.11	0.76, p = 0.33	0.09, p = 0.57	0.20, p = 0.57
		pMLE-DX*	**-0.25**, p = 0.00	**-0.43**, p = 0.00	0.30, p = 0.05	1.7, p = 0.44	0.32, p = 0.35
rs10983712	ASTN2 | TLR4	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.4**, p = 0.00	0.02, p = 0.59	-0.04, p = 0.44
		pMLE-DX	-1.4, p = 0.16	**-4.8**, p = 0.03	1.8, p = 0.27	0.09, p = 0.62	-1.0, p = 0.39
		pMLE-DX*	**-0.25**, p = 0.00	**-0.43**, p = 0.00	1.3, p = 0.10	3.0, p = 0.36	0.009, p = 0.49
rs10983736	ASTN2 | TLR4	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.1**, p = 0.01	-0.24, p = 0.27	0.26, p = 0.26
		pMLE-DX	-1.5, p = 0.14	-2.8, p = 0.09	-0.00, p = 0.67	-0.67, p = 0.34	0.95, p = 0.65
		pMLE-DX*	**-0.25**, p = 0.00	**-0.43**, p = 0.00	**2.8**, p = 0.00	**5.0**, p = 0.038	0.05, p = 0.50
rs16905962	ASTN2 | TLR4	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	-0.005, p = 0.56	-0.14, p = 0.48	1.4,p = 0.15
		pMLE-DX	-1.4, p = 0.23	-2.5, p = 0.15	-0.51, p = 0.51	**5.1**, p = 0.03	1.4, p = 0.72
		pMLE-DX*	**-0.24**, p = 0.00	**-0.43**, p = 0.00	-1.5, p = 0.06	5.0, p = 0.57	3.1, p = 0.49
Rs1927914	ASTN2 | TLR4	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.4**, p = 0.00	0.17, p = 0.12	0.04, p = 0.58
		pMLE-DX	-1.3, p = 0.15	-2.1, p = 0.14	1.2, p = 0.25	0.65, p = 0.76	-0.29, p = 0.47
		pMLE-DX*	**-0.25**, p = 0.00	**-0.43**, p = 0.00	1.2, p = 0.12	3, p = 0.33	1.2, p = 0.08
rs11536879	TLR4	uLR	**-0.25**, p = 0.00	**-0.43**, p = 0.00	0.63, p = 0.25	-0.56, p = 0.30	0.78, p = 0.27
		pMLE-DX	-1.5, p = 0.07	-3, p = 0.056	-2.9, p = 0.37	-3.6, p = 0.36	3.8, p = 0.25
		pMLE-DX*	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**-0.28**, p = 0.04	5, p = 0.19	0.85, p = 0.19
rs4986790	TLR4	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**3.5**, p = 0.02	0.14, p = 0.36	**-3.3**, p = 0.02
		pMLE-DX	-1.4, p = 0.08	-2.4, p = 0.06	**4.9**, p = 0.01	-0.55, p = 0.48	-4.1, p = 0.20
		pMLE-DX*	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**0.84**, p = 0.00	5.0, p = 0.08	-1.3, p = 0.26
rs7045953	TLR4 | LOC100129489	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**0.96**, p = 0.01	-0.02, p = 0.46	0.47, p = 0.11
		pMLE-DX	-1.4, p = 0.10	-2.4, p = 0.09	-0.19, p = 0.54	1.6, p = 0.19	1.1, p = 0.67
		pMLE-DX*	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**2.6**, p = 0.00	4.8, p = 0.47	0.06, p = 0.44
rs7357627	TLR4 | LOC100129489	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.3**, p = 0.00	-0.02, p = 0.43	0.07, p = 0.63
		pMLE-DX	-1.7, p = 0.11	-3.3, p = 0.10	1.1, p = 0.28	-0.31, p = 0.38	-0.24, p = 0.43
		pMLE-DX*	**-0.24**, p = 0.00	**-0.41**, p = 0.00	0.51, p = 0.67	2.3, p = 0.09	1.7, p = 0.08
rs7046020	TLR4 | LOC100129489	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.5**, p = 0.00	0.08, p = 0.69	-0.11, p = 0.33
		pMLE-DX	-1.7, p = 0.14	-2.8, p = 0.11	1.7, p = 0.75	1.1, p = 0.24	-0.91, 0.36
		pMLE-DX*	**-0.25**, p = 0.00	**-0.42**, p = 0.00	**-4.2**, p = 0.00	4.5, p = 0.10	**4.6**, p = 0.01
rs1927937	TLR4 | LOC100129489	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.3**, p = 0.00	-0.02, p = 0.45	0.08, p = 0.63
		pMLE-DX	-1.5, p = 0.11	-2.7, p = 0.09	0.85, p = 0.59	0.25, p = 0.66	0.05, p = 0.50
		pMLE-DX*	-0.25, p = 0.00	-0.42, p = 0.00	1.3, p = 0.07	4.3, p = 0.18	-0.51, p = 0.20
rs1927924	TLR4 | LOC100129489	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.6**, p = 0.00	0.25, p = 0.18	-0.20, p = 0.28
		pMLE-DX	-1.5, p = 0.23	-2.9, p = 0.21	**4.9**, p = 0.001	**4.3**, p = 0.05	-4.2, p = 0.17
		pMLE-DX*	**-0.24**, p = 0.00	**-0.43**, p = 0.00	0.15, p = 0.61	5.0, p = 0.25	**1.6**, p = 0.04
rs3860141	TLR4 | LOC100129489	uLR	**-0.24**, p = 0.00	**-0.42**, p = 0.00	**1.1**, p = 0.00	-0.06, p = 0.33	**0.37**, p = 0.02
		pMLE-DX pMLE-DX*	-1.2, p = 0.16	-2.2, p = 0.11	0.01, p = 0.55	-0.35, p = 0.35	1.1, p = 0.69
			-0.24, p = 0.00	**-0.42**, p = 0.00	1.4, p = 0.17	3.6, p = 0.53	-0.57, p = 0.46
rs1877876	TLR4 | LOC100129489	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.7**, p = 0.00	0.20, p = 0.11	-0.31, p = 0.09
		pMLE-DX pMLE-DX*	-1.5, p = 0.10	-2.3, p = 0.08	2.3, p = 0.30	1.1, p = 0.22	-1.5, p = 0.32
			-0.24, p = 0.00	**-0.42**, p = 0.00	-1.2, p = 0.07	3.9, p = 0.56	**2.3**, p = 0.02
rs497322	TLR4 | LOC100129489	uLR	**-0.25**, p = 0.00	**-0.43**, p = 0.00	**1.7**, p = 0.01	**0.69**, p = 0.01	-0.28, p = 0.28
		pMLE-DX pMLE-DX*	-1.4, p = 0.23	-1.9, p = 0.20	1.8, p = 0.26	**4.5**, p = 0.04	-1, p = 0.42
			-0.24, p = 0.00	**-0.43**, p = 0.00	**0.24**, p = 0.03	**5.0**, p = 0.05	1.7, p = 0.47
rs6478330	TLR4 | LOC100129489	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.2**, p = 0.03	0.36, p = 0.21	0.16, p = 0.28
		pMLE-DX pMLE-DX*	-1.5, p = 0.15	-2.6, p = 0.13	-0.008, p = 0.62	0.73, p = 0.24	0.93, p = 0.33
			-0.25, p = 0.00	**-0.43**, p = 0.00	**-2.8**, p = 0.048	5.0, p = 0.14	**5.5**, p = 0.01
rs7856175	TLR4 | LOC100129489	uLR	**-0.24**, p = 0.00	**-0.43**, p = 0.00	**1.7**, p = 0.00	0.07, p = 0.30	**-0.33**, p = 0.03
		pMLE-DX pMLE-DX*	-1.4, p = 0.13	-2.3, p = 0.11	1.6, p = 0.24	0.43, p = 0.28	-0.97, p = 0.36
			-0.25, p = 0.00	**-0.43**, p = 0.00	**1.5**, p = 0.03	2.7, p = 0.61	-0.08, p = 0.28
rs3134940	AGER	uLR	**-0.97,** p = 0.00	-0.12, p = 0.08	**1.1**, p = 0.00	**0.55**, p = 0.00	-0.28, p = 0.05
		pMLE-DX	-0.74, p = 0.13	-0.52, p = 0.14	1.2, p = 0.40	1.2, p = 0.18	-0.67, p = 0.33
		pMLE-DX*	-0.96, p = 0.11	-1.2, p = 0.10	1.4, p = 0.09	1.8, p = 0.13	-0.32, p = 0.50

(*Continued*)

Table 5. (Continued)

SNP	Gene/Intergenic Region	Method	$\hat{\beta}_{Age}$	$\hat{\beta}_{Sex}$	$\hat{\beta}_{\varepsilon 4}$	$\hat{\beta}_{G}$	$\hat{\beta}_{G\times\varepsilon 4}$
rs1035798	AGER	uLR	**-0.97**, p = 0.00	**-0.13**, p = 0.06	**0.89**, p = 0.00	**0.43**, p = 0.00	0.03, p = 0.85
		pMLE-DX	-1.42, p = 0.09	-0.29, p = 0.18	0.96, p = 0.56	0.67, p = 0.21	-0.06, p = 0.41
		pMLE-DX*	-0.97, p = 0.11	-0.60, p = 0.15	1.3, p = 0.08	1.8, p = 0.13	-1.9, p = 0.08
rs2070600	AGER	uLR	**-0.97**, p = 0.00	-0.12, p = 0.09	**0.99**, p = 0.00	**0.49**, p = 0.00	**-0.14**, p = 0.33
		pMLE-DX	-2.3, p = 0.08	-0.50, p = 0.18	0.97, p = 0.59	0.99, p = 0.22	-0.24, p = 0.17
		pMLE-DX*	-0.97, p = 0.11	-0.62, p = 0.13	1.5, p = 0.05	1.8, p = 0.13	-0.23, p = 0.51

Analyses are performed using the usual logistic regression (uLR) that uses the clinical diagnosis as an outcome and using pseudo-likelihood method that assumes that the proportion of nuisance disease within the clinically diagnosed AD is 36% for $\varepsilon 4$ non-carriers and is 6% for $\varepsilon 4$ carriers. Pseudo-likelihood analyses pMLE-DX estimates parameters for $D = 1$ vs. $D = 0$ and $D = 1^*$ combined. Pseudo-likelihood analyses pMLE – DX*, however, estimate two sets of risk coefficients, i.e. βs for $D = 0$ vs. $D = 1$ and β^*s $D = 0$ vs. $D = 1^*$. Estimates of β^*s are reported in **S7 Table**.

https://doi.org/10.1371/journal.pone.0201140.t005

One SNP, rs830832, has significant $\hat{\beta}_{G\times\varepsilon 4}$ both in uLR ($\hat{\beta}_{G\times\varepsilon 4} = 0.74, p = 0.01$) and pMLE $-$ DX* ($\hat{\beta}_{(G\times\varepsilon 4)} = 2.6, p = 0.03$). This SNP locates at the intergenic region between SORBS2 and TLR3 at Chromosome 4 and are 72k downstream of SNP rs75718659, which was reported associated with Alzheimer's disease in a family-based GWAS [10].

Among the seven SNPs appear to have significant $\hat{\beta}_{G\times\varepsilon 4}$ in pMLE- DX* but not uLR, two of the SNPs: rs4862611 ($\hat{\beta}_{G\times\varepsilon 4}$ = -2.8, p = 0.03) and rs1706143 ($\hat{\beta}_{G\times\varepsilon 4}$ = 2.9, p = 0.03), are also located at the intergenic region between SORBS2 and TLR3 at Chromosome 4 and are 80k and 20k downstream of SNP rs75718659.

Nine of the SNPs appear to have significant $\hat{\beta}^*_{G\times\varepsilon 4}$ in pMLE-DX* but not $\hat{\beta}_{G\times\varepsilon 4}$ in uLR or pMLE-DX. Two SNPs, rs7676342 ($\hat{\beta}^*_{G\times\varepsilon 4} = -2.1, p = 0.02$) and rs13113778 ($\hat{\beta}^*_{G\times\varepsilon 4} = -2.7$, $p = 0.03$), again are located in the intergenic region between SORBS2 and TLR3 at Chromosome 4 and are 80k and 100k downstream of SNP rs75718659, respectively. Three SNPs, rs955302 ($\hat{\beta}^*_{G\times\varepsilon 4} = 4.0, p = 0.01$), rs4837254 ($\hat{\beta}^*_{G\times\varepsilon 4} = 2.3, p = 0.04$) and rs12342331 ($\hat{\beta}^*_{G\times\varepsilon 4} = 2.1, p = 0.04$), are located at the intergenic region between ASTN2 and TLR4 at Chromosome 9 and are 400k, 430k, and 492k downstream of rs1360695 associated with Schizophrenia [11].

Estimates of β_G, however, are generally larger in magnitude when estimated in pMLE-DX and pMLE- DX* models.

Two SNPs appear to be associated with the diagnosis both in uLR and pMLE-DX. SNP rs7656500 (uLR $\hat{\beta}_G = 0.79, p = 0.01$ and pMLE-DX $\hat{\beta}_G = 6, p = 0.02$) locates at the intergenic region between KIAA0922 and TLR2 at Chromosome 4, and is 163k upstream and 144k downstream of rs727153 and rs1466662, respectively, which were reported associated with Alzheimer's disease in two studies [12,13]. It is also 54k upstream of rs7654093 associated with thrombosis [14], 30k upstream of rs7659024 associated with Venous thromboembolism [15], 34k upstream of rs2066865 associated with Venous thromboembolism [16, 17], 52k upstream of rs6536024 associated with Venous thromboembolism [18], and 360k downstream of rs11099942 associated with Type 2 diabetes [19].

Among six SNPs which appear to be significantly associated with the nuisance diagnosis in absence of an interactive effect, three SNPs rs1869617 (at the intergenic region between SORBS2 and TLR3 at Chromosome 4, pMLE- DX* $\hat{\beta}^*_G = 4.9, p = 0.01$), rs3775296 (at the UTR region of TLR3 at Chromosome 4, pMLE- DX* $\hat{\beta}^*_G = 4.3, p = 0.046$), rs7668666 (at the INTRON region of TLR3 at Chromosome 4, pMLE- DX* $\hat{\beta}^*_G = 4.9, p = 0.00$) locate 110k,

176k and 179k, respectively, downstream of rs75718659 reported associated with Alzheimer's disease [10] and another two SNPs rs16905625 (pMLE- DX* $\hat{\beta}^*_G = 3.2, p = 0.03$) and rs1890047 (pMLE- DX* $\hat{\beta}^*_G = 4.6, p = 0.008$) locate at the intergenic region between ASTN2 and TLR4 at Chromosome 9, 412k and 428k, respectively downstream of rs1360695 reported associated with Schizophrenia [11].

Estimates of $\beta_{\varepsilon 4}$ in the absence of interaction are generally larger in magnitude for the diagnosis of interest in pMLE-DX. For example, in a model with SNP rs1816702 (uLR $\hat{\beta}_{\varepsilon 4} = 1.4, p = 0.00$ and pMLE- DX* $\hat{\beta}_{\varepsilon 4} = 2.6, p = 0.03, \hat{\beta}^*_{\varepsilon 4} = 1.2, p = 0.01$).

AGER. All of the three SNPs in the AGER gene measured in the data are associated with susceptibility to AD as inferred in uLR and also are associated with susceptibility to the nuisance disease when measured by pMLE- DX*. **rs3134940** has been previously reported in association to breast cancer, type I diabetes and other phenotypes (https://www.gwascentral.org/marker/HGVM1600838/results?t=ZERO); **rs1035798** and **rs2070600** have been previously reported in association to rheumatoid arthritis (https://www.gwascentral.org/marker/HGVM275161/results?t=ZERO and https://www.gwascentral.org/marker/HGVM571318/results?t=ZERO).

Discussion

We investigated if disease heterogeneity among clinically diagnosed cases could introduce bias into the estimates of GxE interactions. We showed that when there is a strong association between the environmental variable and the relative risk of the disease of interest, as compared to the nuisance disease, and then there could be bias in either direction. We base our developments on the method by Chatterjee and Carroll [7] that is fully efficient in situations when the genetic and environmental variables are distributed independently in the population, a population-based genetics model is assumed for the genetic factors and the environmental variables are treated non-parametrically.

Interestingly, in our analyses, the estimates of regression coefficients are qualitatively differed between the analyses that used the clinical diagnosis as a surrogate of the pathologic diagnosis and the analyses that used our newly proposed pseudo-likelihood approach that incorporates the uncertainty of the clinical diagnosis. Specifically, in TLR set for 13% of the SNPs examined, GxE was found to be significant in the relationship to the clinical diagnosis, while the pseudo-likelihood analyses inferred these GxE to be not significant. On the other hand, for 14% of the SNPs that we examined, GxE was found to be statistically significant only when we incorporated the uncertainty in the clinical-pathological diagnoses relationship. This finding is consistent with the conclusion reached by a study of phenotypic misclassification among cases [20] in situations when the misclassification is non-differential, i.e. is not a function of the environmental variables. The study concluded that presence of "non-cases" greatly decreased the estimates of risk attributed to the genetic variation.

One of the major concerns in the analyses of the genetic studies has been the missing heritability, when the genetic markers identified thus far explain only a small portion of inter-person variability in familiar clustering of complex diseases [21]. The downward biases in the estimates associating GxE to the clinically diagnosed disease status might in part explain the missing heritability. On the other hand, the upward biases in these estimates might in part address the conclusion reached by [22] that only 1% of the association found are likely to be true.

We examined estimates of the genetic effects, ApoE4 status, and age, sex consistent with the original publication on this dataset [23]. Epidemiologic evidence [24] suggests that the following factors play important role in AD risk: education/cognitive reserve, racial and ethnic

difference, gender, smoking, drinking, head injury, diabetes, cardiovascular disease, obesity, social engagement, etc. However, not all of these factors have been consistently confirmed by subsequent studies, and considerable inconsistencies exist. For example, nicotine intake has been observed to decrease the risk of dementia due to the demonstrated ability of nicotine to stimulate neurotransmitter systems that are compromised in dementia [25]. More recent studies have suggested that nicotine intake may increase the risk of AD and also bring forward age of onset with APOE interactive effect [26].

The main conclusion reached in this paper is that using the clinically diagnosed status can lead to severely biased estimates of GxE interactions in situations when the frequency of the pathologic diagnosis of interest, as compared to other diagnoses, depends on the environment, and we aim to correct such biases by proposing pseudolikelihood method. AD dataset is mainly used for illustration, therefore, for clarity we restricted to variables to the minimum necessary instead of considering full risk prediction modes which might be able to better describe the inter-patient variability in susceptibility to AD. Although other factors are potentially important in predicting the risk of AD, this relatively simple model was able to achieve the main goals of the current manuscript. By recognizing and accounting for the potential of case heterogeneity, which biases the gene x environment interaction, our newly proposed method has the ability to remove this bias.

Define E to be the set of variables in the model, i.e. age, sex. Let O define a set of key environmental variables omitted from the model. Addition of variables O would not modify the effect estimates of GxE beyond what is expected purely by chance if O does not interact with either G or E. Also, if conditional on the diagnosis of AD, GxE is independent of O, then omission of O does not change the effect estimate of GxE [27]. If, however, O interacts with GxE, then addition of these variables would change the effect estimate of GxE in the direction that is consistent with the direction of the GxE effect. Further studies that incorporate environmental variables, such as medical history, tobacco use, and infections are needed for their potential to modify the risk and the estimates of GxE in particular.

Epigenetic mechanisms are well-recognized in the mediation of GxE and analysis of epigenetic changes at the genome scale can offer new insights into the relationship between brain epigenomes and AD. Further, candidate genes from epigenome-wide association studies interact with those from GWAS that can undergo epigenetic changes in their upstream gene regulatory elements [28]. However, an active conundrum is how the epigenetic mechanisms influence gene-environment interactions.

Appendix

Derivation of pseudo-likelihood (2) and covariance matrix

Derivation of the pseudo-likelihood (2) is straightforward.

Next we demonstrate that the pseudo-likelihood (2) has zero mean when evaluated at the true parameters. Derivative of (2) with respect to Ω is

$$\frac{\sum_{d^{\blacksquare}} \gamma_{d^{\blacksquare}|d^{cl}}(x) \times \mathrm{S}_{\Omega}(d^{\blacksquare}, d^{cl}, g, x, z; \Omega)}{\sum_{d^{\blacksquare}} \gamma_{d^{\blacksquare}|d^{cl}}(x) \times \mathrm{S}(d^{\blacksquare}, d^{cl}, g, x, z; \Omega)} - \frac{\sum_{d^{\blacksquare},g^{\blacksquare},d^{cl\blacksquare}} \gamma_{d^{\blacksquare}|d^{cl\blacksquare}}(x) \times \mathrm{S}_{\Omega}(d^{\blacksquare}, d^{cl}, g^{\blacksquare}, x, z; \Omega)}{\sum_{d^{\blacksquare},g^{\blacksquare},d^{cl\blacksquare}} \gamma_{d^{\blacksquare}|d^{cl\blacksquare}}(x) \times \mathrm{S}(d^{\blacksquare}, d^{cl}, g^{\blacksquare}, x, z; \Omega)}$$
$$= A(d^{cl}, g, x, z) - B(x, z).$$

Let $\mathrm{p}(x,z|\eta)$ be the density of the environment.

Note the conditional probabilities

$$[G, X, Z|D^{CL}] = n_{d^{cl}}^{-1} \sum_{d^{\blacksquare}} \gamma_{d^{cl}|d^{\blacksquare}}(x) \times S(d^{\blacksquare}, d^{cl}, g, x, z; \Omega),$$

$$[X, Z|D^{CL}] = n_{d^{cl}}^{-1} \sum_{g^{\blacksquare}, d^{\blacksquare}} \gamma_{d^{cl}|d^{\blacksquare}}(x) \times S(d^{\blacksquare}, d^{cl}, g^{\blacksquare}, x, z; \Omega) \times p(x, z|\eta).$$

Hence

$$E\{A(D^{CL}, G, X, Z)\} = \sum_{d^{cs}} \frac{n_{d^{cl}\blacksquare}}{n} E\{A(D^{CL}, G, X, Z)|D^{CL} = d^{cl\blacksquare}\}$$

$$= \frac{1}{n} \sum_{d^{\blacksquare}, d^{cl\blacksquare}, g^{\blacksquare}, x^{\blacksquare}, z^{\blacksquare}} \gamma_{d^{cl}|d^{\blacksquare}}(x^{\blacksquare}) \times \mathrm{S}_{\Omega}(d^{\blacksquare}, d^{cl\blacksquare}, g^{\blacksquare}, x^{\blacksquare}, z^{\blacksquare}; \Omega) \times p(x^{\blacksquare}, z^{\blacksquare}|\eta)$$

$$= \sum_{d^{cl\blacksquare}} \frac{n_{d^{cl}\blacksquare}}{n} E\{B(X, Z)|D^{CL} = d^{cl\blacksquare}\} = E\{B(X, Z)\}.$$

Therefore the derivative of the pseudo-likelihood has zero mean when evaluated at the true parameters. Evaluated at the true parameters the estimating function (2) takes the following form

$$n^{-1/2} \sum_{i=1}^{n} E[A(d^{cl}, g, x, z) - B(x, z) - E\{A(d^{cl}, g, x, z) - B(x, z)|D^{CL} = d^{cl}\}].$$

Covariance matrix is then

$$\Sigma = n^{-1} \sum_{i=1}^{n} E[\{A(d^{cl}, g, x, z) - B(x, z)\} \times \{A(d^{cl}, g, x, z) - B(x, z)\}^{T}] - \Lambda.$$

Define

$$Q_1(d^{cl}, g, x, z; \Omega) = \sum_{d^{\blacksquare}} \gamma_{d^{cl}|d^{\blacksquare}}(x) \times S_{\Omega}(d^{\blacksquare}, d^{cl}, g, x, z; \Omega) \times \mathrm{p}(x, z|\eta);$$

$$Q_2(d^{cl}, g, x, z; \Omega) = \sum_{d^{\blacksquare}} \gamma_{d^{cl}|d^{\blacksquare}}(x) \times S(d^{\blacksquare}, d^{cl}, g, x, z; \Omega) \times \mathrm{p}(x, z|\eta);$$

$$Q_3(x, z; \Omega) = \sum_{d^{cl\blacksquare}, d^{\blacksquare}, g^{\blacksquare}} \gamma_{d^{cl}|d^{\blacksquare}}(x) \times S_{\Omega}(d^{\blacksquare}, d^{cl\blacksquare}, g^{\blacksquare}, x, z; \Omega) \times \mathrm{p}(x, z|\eta);$$

$$Q_4(x, z; \Omega) = \sum_{d^{cl\blacksquare}, d^{\blacksquare}, g^{\blacksquare}} \gamma_{d^{cl}|d^{\blacksquare}}(x) \times S(d^{\blacksquare}, d^{cl\blacksquare}, g^{\blacksquare}, x, z; \Omega) \times \mathrm{p}(x, z|\eta);$$

$$\Sigma_1 = n^{-1} \sum_{d^{cl\blacksquare}, g^{\blacksquare}, x^{\blacksquare}, z^{\blacksquare}} \frac{Q_1(d^{cl\blacksquare}, g^{\blacksquare}, x^{\blacksquare}, z^{\blacksquare}; \Omega) \times Q_1^{T}(d^{cl\blacksquare}, g^{\blacksquare}, x^{\blacksquare}, z^{\blacksquare}; \Omega)}{Q_2(d^{cl\blacksquare}, g^{\blacksquare}, x^{\blacksquare}, z^{\blacksquare}; \Omega)} p(x^{\blacksquare}, z^{\blacksquare}|\eta);$$

$$\Sigma_2 = n^{-1} \sum_{x^{\blacksquare}, z^{\blacksquare}} \frac{Q_3(x^{\blacksquare}, z^{\blacksquare}; \Omega) \times Q_3^{T}(x^{\blacksquare}, z^{\blacksquare}; \Omega)}{Q_4(x^{\blacksquare}, z^{\blacksquare}; \Omega)} p(x^{\blacksquare}, z^{\blacksquare}|\eta).$$

The covariance matrix can then be represented in the form $\Sigma = \Sigma_1 - \Sigma_2 - \Lambda$.

Define $I_1 = \sum_{d^{cl\blacksquare}} \frac{n_{d^{cl}\blacksquare}}{n} E\left[\frac{\partial}{\partial \Omega} \left\{\frac{Q_1(d^{cl\blacksquare}, g, x, z; \Omega)}{Q_2(d^{cl\blacksquare}, g, x, z; \Omega)}|D^{CL} = d^{cl}\right\}\right]$ and $I_2 = \sum_{d^{cl\blacksquare}} \frac{n_{d^{cl}\blacksquare}}{n} E\left[\frac{\partial}{\partial \Omega} \left\{\frac{Q_3(d^{cl\blacksquare}, g, x, z; \Omega)}{Q_4(d^{cl\blacksquare}, g, x, z; \Omega)}|D^{CL} = d^{cl}\right\}\right]$, then $I = I_1 - I_2$.

We note that $I_1 = n^{-1}\frac{\partial^2}{\partial\Omega\partial\Omega^T}\sum_{d^{cl\blacksquare},d^{\blacksquare},g^{\blacksquare},x^{\blacksquare},z^{\blacksquare}}\gamma_{d^{cl}|d^{\blacksquare}}(x^{\blacksquare}) \times S_\Omega(d^{\blacksquare}, d^{cl\blacksquare}, g^{\blacksquare}, x^{\blacksquare}, z^{\blacksquare}; \Omega) \times \mathrm{p}(x^{\blacksquare}, z^{\blacksquare}|\eta) + \Sigma_1$ and $I_2 = n^{-1}\frac{\partial^2}{\partial\Omega\partial\Omega^T}\sum_{d^{cl\blacksquare},d^{\blacksquare},g^{\blacksquare},x^{\blacksquare},z^{\blacksquare}}\gamma_{d^{cl}|d^{\blacksquare}}(x^{\blacksquare}) \times S_\Omega(d^{\blacksquare}, d^{cl\blacksquare}, g^{\blacksquare}, x^{\blacksquare}, z^{\blacksquare}; \Omega) \times \mathrm{p}(x^{\blacksquare}, z^{\blacksquare}|\eta) + \Sigma_2$. Hence $\Sigma = I_1 - I_2 - \Lambda = I - \Lambda$.

Supporting information

S1 Table. $\boldsymbol{\beta_{G\times\varepsilon 4} \neq 0}$. The Bias and Root Mean Squared Error (RMSE) in parameter estimates from simulations using the usual logistic regression with clinical diagnosis as the outcome (uLR), the pseudo-likelihood approach (pMLE), and our newly proposed pseudo-likelihood approach that accounts for misdiagnosis (pMLE-DX). For these simulations, the study included n_0 controls and n_1 cases. Frequency of ApoE ε4 allele in the population is 14%. Variables Z_1 and Z_2 are Bernoulli with frequencies 0.50 and 0.52, respectively. Frequency of the *true* disease status is 46% in the population; and is 40% among the subpopulation with no ApoE ε4 alleles, and 82% in the subpopulation with at least one ApoE ε4 alleles. Frequency of nuisance disease within the clinical diagnosis varies by ApoE4 status $\mathrm{pr}(\mathrm{D} = 1^*|D^{CL} = 1,\varepsilon 4-) = 0.36$ and $\mathrm{pr}(\mathrm{D} = 1^*|D^{CL} = 1,\varepsilon 4) = 0.06$.
(DOCX)

S2 Table. $\boldsymbol{\beta_{G\times\varepsilon 4} = 0}$. The Bias and Root Mean Squared Error (RMSE) in parameter estimates from simulations using the usual logistic regression with clinical diagnosis as the outcome (uLR), the pseudo-likelihood approach (pMLE), and our newly proposed pseudo-likelihood approach that accounts for misdiagnosis (pMLE-DX). For these simulations, the study included n_0 controls and n_1 cases. Frequency of ApoE ε4 allele in the population is 14%. Variables Z_1 and Z_2 are Bernoulli with frequencies 0.50 and 0.52, respectively. Frequency of the *true* disease status is 46% in the population; and is 40% among the subpopulation with no ApoE ε4 alleles, and 82% in the subpopulation with at least one ApoE ε4 alleles. Frequency of nuisance disease within the clinical diagnosis varies by ApoE4 status $\mathrm{pr}(\mathrm{D} = 1^*|D^{CL} = 1,\varepsilon 4-) = 0.36$ and $\mathrm{pr}(\mathrm{D} = 1^*|D^{CL} = 1,\varepsilon 4+) = 0.06$.
(DOCX)

S3 Table. Frequency of the nuisance disease is underestimated. The Bias and Root Mean Squared Error (RMSE) in parameter estimates from simulations using the usual logistic regression with clinical diagnosis as the outcome (uLR), the pseudo-likelihood approach (pMLE), and our newly proposed pseudo-likelihood approach that accounts for misdiagnosis (pMLE-DX). For these simulations, the study included $n_0 = 3000$ controls and $n_1 = 3000$ cases. Frequency of ApoE ε4 allele in the population is 14%. Variables Z_1 and Z_2 are Bernoulli with frequencies 0.50 and 0.52, respectively. Frequency of the *true* disease status is 46% in the population; and is 40% among the subpopulation with no ApoE ε4 alleles, and 82% in the subpopulation with at least one ApoE ε4 alleles. Frequency of nuisance disease within the clinical diagnosis varies by ApoE4 status $\mathrm{pr}(\mathrm{D} = 1^*|D^{CL} = 1,\varepsilon 4-) = 0.36$ and $\mathrm{pr}(\mathrm{D} = 1^*|D^{CL} = 1,\varepsilon 4+) = 0.06$. The clinical-pathophysiological diagnoses relationship is misspecified to be $\mathrm{pr}(\mathrm{D} = 1^*|D^{CL} = 1,\varepsilon 4-) = 0.30$ and $\mathrm{pr}(\mathrm{D} = 1^*|D^{CL} = 1,\varepsilon 4+) = 0$.
(DOCX)

S4 Table. Frequency of the nuisance disease is overestimated. Bias and Root Mean Squared Error (RMSE) for parameter estimates based on a study of 500 simulated datasets with n_0 controls and n_1 cases with clinical phenotype. Analyses are based on the usual logistic regression model that ignores nuisance disease and based on pseudolikelihood with (pMLE-DX) and without the consideration of clinical-pathological diagnoses relationship (pMLE). Frequency of ApoE ε4 alleles is 14% in the population. Variables Z_1 and Z_2 are Bernoulli with frequencies

0.50 and 0.52, respectively. Frequency of the *true* disease status is 46% in the population; and is 40% among the subpopulation with no ApoE $\varepsilon 4$ alleles, and 82% in the subpopulation with at least one ApoE $\varepsilon 4$ alleles. Frequency of nuisance disease within the clinical diagnosis varies by ApoE4 status pr(D = 1′|D^{CL} = 1,ε4−) = 0.36 and pr(D = 1′|D^{CL} = 1,ε4+) = 0.06. The clinical-pathological diagnoses relationship is misspecified to be pr(D = 1′|D^{CL} = 1,ε4−) = 0.42 and pr (D = 1′|D^{CL} = 1,ε4+) = 0.12.
(DOCX)

S5 Table. $\boldsymbol{\beta^*_{G\times\varepsilon 4} = 0}$ and $\boldsymbol{\beta^*_G = 0}$. The Bias and Root Mean Squared Error (RMSE) in parameter estimates from simulations using the usual logistic regression with clinical diagnosis as t he outcome (uLR), the pseudo-likelihood approach (pMLE), and our newly proposed pseudo-likelihood approach that accounts for misdiagnosis (pMLE-DX). For these simulations, the study included n_0 controls and n_1 cases. Risk of the disease of interest is defined in a set of parameters $\beta_0, \beta_G, \beta_{Z_1}, \beta_{Z_2}, \beta_{G\times\varepsilon 4}$; while the risk of the nuisance disease is parametrized by $\beta^*_0, \beta^*_G, \beta^*_{\varepsilon 4}, \beta^*_{G\times\varepsilon 4}$. Frequency of ApoE $\varepsilon 4$ allele in the population is 14%. Variables Z_1 and Z_2 are Bernoulli with frequencies 0.50 and 0.52, respectively. Frequencies of the disease of interest and the nuisance disease are pr(D = 1) = 24.8%, pr(D = 1*) = 12.5%, pr(D = 1|ε4+) = 43%, pr (D = 1*|ε4+) = 16.1%, pr(D = 1|ε4−) = 20%, pr(D = 1*|ε4−) = 11.6%. Frequency of the nuisance disease within the clinical diagnosis varies by ApoE4 status pr(D = 1*|D^{CL} = 1,ε4−) = 0.36 and pr(D = 1*|D^{CL} = 1,ε4+) = 0.06.
(DOCX)

S6 Table. $\boldsymbol{\beta_{G\times\varepsilon 4} = 0, \beta^*_{G\times\varepsilon 4} = 0, \beta^*_G = 0}$. The Bias and Root Mean Squared Error (RMSE) in parameter estimates from simulations using the usual logistic regression with clinical diagnosis as the outcome (uLR), the pseudo-likelihood approach (pMLE), and our newly proposed pseudo-likelihood approach that accounts for misdiagnosis (pMLE-DX). For these simulations, the study included n_0 controls and n_1 cases. Risk of the disease of interest is defined in a set of parameters $\beta_0, \beta_G, \beta_{Z_1}, \beta_{Z_2}, \beta_{G\times\varepsilon 4}$; while the risk of the nuisance disease is parametrized by $\beta^*_0, \beta^*_G, \beta^*_{\varepsilon 4}, \beta^*_{G\times\varepsilon 4}$. Frequency of ApoE $\varepsilon 4$ allele in the population is 14%. Variables Z_1 and Z_2 are Bernoulli with frequencies 0.50 and 0.52, respectively. Frequencies of the disease of interest and the nuisance disease are pr(D = 1) = 24.8%, pr(D = 1*) = 12.5%, pr(D = 1|ε4+) = 43%, pr (D = 1*|*ApoE*4+) = 16.1%, pr(D = 1|ε4−) = 20%, pr(D = 1*|ε4−) = 11.6%. Frequency of the nuisance disease within the clinical diagnosis varies by ApoE4 status pr(D = 1*|D^{CL} = 1,ε4−) = 0.36 and pr(D = 1*|D^{CL} = 1,ε4+) = 0.06.
(DOCX)

S7 Table. Parameter estimates in Alzheimer's disease study. Analyses are performed using the usual logistic regression (uLR) that uses the clinical diagnosis as an outcome and using pseudo-likelihood method that assumes that the proportion of nuisance disease within the clinically diagnosed AD is 36% for $\varepsilon 4$ carriers and is 6% for $\varepsilon 4$ non-carriers. Pseudo-likelihood analyses pMLE-DX estimates parameters for $D = 1$ vs. $D = 0$ and $D = 1^*$ combined. Pseudo-likelihood analyses pMLE − DX*, however, estimate two sets of risk coefficients, i.e. βs for $D = 0$ vs. $D = 1$ and β^*s $D = 0$ vs. $D = 1^*$.
(DOCX)

S8 Table. SNPs previously reported in GWAS that are within 500k up- or downstream of SNPs that we inferred in Alzheimer's disease study. (Table 5, SNPs whose effect estimates of β_G and/or $\beta_{G\times\varepsilon 4}$ are with permutation-based p-value <0.05).
(DOCX)

Acknowledgments

Dr. Lobach and Dr. Zhang are supported by 5R21AG043710-02.

The Alzheimer's disease dataset is available in the Database of Genotypes and Phenotypes study accession number phs000372.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000372.v1.p1).

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).

Author Contributions

Conceptualization: Iryna Lobach, Joshua Sampson.

Data curation: Li Zhang.

Formal analysis: Iryna Lobach, Joshua Sampson, Alexander Alekseyenko, Siarhei Lobach, Li Zhang.

Funding acquisition: Iryna Lobach.

Investigation: Iryna Lobach, Li Zhang.

Methodology: Iryna Lobach, Joshua Sampson, Siarhei Lobach.

Resources: Iryna Lobach.

Software: Iryna Lobach.

Validation: Iryna Lobach.

Writing – original draft: Iryna Lobach.

Writing – review & editing: Iryna Lobach, Joshua Sampson, Alexander Alekseyenko, Siarhei Lobach, Li Zhang.

References

1. Potter H, Wisniewski T. Apolipoprotein E: essential catalyst of the Alzheimer amyloid cascade. International Journal of Alzheimer's Disease. 2012; http://dx.doi.org/10.1155/2012/489428
2. Salloway S, Sperling R. Understanding conflicting neurological findings in patients clinically diagnosed as having Alzheimer Dementia. JAMA Neurology. 2015; 72 (10): 1106–1108. https://doi.org/10.1001/jamaneurol.2015.1804 PMID: 26302229
3. Shaw AC, Panda A, Joshi SR, Qian F, Allore HG, Montgomery RR. Dysregulation of human toll-like receptor function in aging. Ageing Research Review. 2011 Jul; 10(3):346–53. https://doi.org/10.1016/j.arr.2010.10.007 PMID: 21074638
4. Ramasamy R, Vannucci SJ, Yan SSD, Herold K, Yan SF, Schmidt AM. Advanced glycation end products and RAGE: a common thread in aging, diabetes, neurodegeneration, and inflammation. Glycobiology. 2005; 15(7): 16R–28R. https://doi.org/10.1093/glycob/cwi053 PMID: 15764591
5. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu. Measurement error in nonlinear models: a modern perspective. 2nd ed. Chapman and Hall/CRC; 2006.

6. Prentice KL, Pyke DA. Logistic disease incidence models and case-control studies, Biometrika.1979; 66(3): 403–411.
7. Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika. 2005; 92(2): 399–418.
8. Hardy GH. Mendelian Proportions in a Mixed Population. Science. 1908; 28(706): 49–50. https://doi.org/10.1126/science.28.706.49 PMID: 17779291
9. Benjamini Y., and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series. 1995; B 57: 289–300.
10. Herold C, Hooli BV, Mullin K, Liu T, Roehr JT, Mattheisen M, et al. (2016) Family-based association analyses of imputed genotypes reveal genome-wide significant association of alzheimer's disease with OSBPL6, PTPRG and PDCL3. Molecular Psychiatry. 2016; 21(11):1608–1612. https://doi.org/10.1038/mp.2015.218 PMID: 26830138
11. Goes FS, McGrath J, Avramopoulos D, Wolyniec P, Pirooznia M, Ruczinki I, et al. (2015) Genome-wide association study of schizophrenia in Ashkenazi Jews. American Journal of Medical Genetics. 2015; 168(8):649–659. https://doi.org/10.1002/ajmg.b.32349 PMID: 26198764
12. Abraham R, Moskvina V, Sims R, Hollingworth P, Morgan A, Georgieva L, et al. A genome-wide association study for late-onset Alzheimer's disease using DNA pooling, BMC Medical Genomics. 2008; 29: 1–44.
13. Kamboh MI, Barmada MM, Demirci FY, Minster RL, Carrasquillo MM, Pankratz VS, et al. Genome-wide association analysis of age-at-onset in Alzheimer's disease. Molecular Psychiatry. 2012 Dec; 17 (12):1340–6. https://doi.org/10.1038/mp.2011.135 PMID: 22005931
14. Hinds DA, Buil A, Ziemek D, Martinez-Perez A, Malik R, Folkersen L et al. Genome-wide association analysis of self-reported events in 6135 individuals and 252 827 controls identifies 8 loci associated with thrombosis. Human Molecular Genetics. 2016; 25(9):1867–1874. https://doi.org/10.1093/hmg/ddw037 PMID: 26908601
15. Germain M, Saut N, Greliche N, Dina C, Lambert JC, Perret C, et al. Genetics of venous thrombosis: insights from a new genome wide association study. PLOS One. 2011; 6(9): e25581 https://doi.org/10.1371/journal.pone.0025581 PMID: 21980494
16. Germain M, Chasman DI, de Haan H, Tang W, Lindström S, Weng LC, et al. Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. American Journal of Human Genetics. 2015 Apr 2; 96(4):532–42. https://doi.org/10.1016/j.ajhg.2015.01.019 PMID: 25772935
17. Klarin D, Emdin CA, Natarajan P, Conrad MF, INVENT Consortium, Kathiresan S. Genetic Analysis of Venous Thromboembolism in UK Biobank Identifies the ZFPM2 Locus and Implicates Obesity as a Causal Risk Factor. Circulation Cardiovascular Genetics. 2017 Apr; 10(2). pii: e001643. https://doi.org/10.1161/CIRCGENETICS.116.001643 PMID: 28373160
18. Tang W, Teichert M, Chasman DI, Heit JA, Morange PE, Li GA genome-wide association study for venous thromboembolism: the extended Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, Genetic Epidemiology Journal 2013 37(5): 512–521
19. Hamet P, Haloui M, Harvey F, Marois-Blanchet FC, Sylvestre MP, Tahir MR, et al. PROX1 gene CC genotype as a major determinant of early onset of type 2 diabetes in slavic study participants from Action in Diabetes and Vascular Disease: Preterax and Diamicron MR Controlled Evaluation study, Journal of Hypertension. 2017 May; 35 Suppl 1:S24–S32. https://doi.org/10.1097/HJH.0000000000001241 PMID: 28060188
20. Manchia M, Cullis J, Gustavo T, Rouleau GY, Uher R, Alda M. The impact of phenotypic and genetic heterogeneity on results of genome-wide association studies of complex diseases. PLOS One. 2013; 8 (10): e76295. https://doi.org/10.1371/journal.pone.0076295 PMID: 24146854
21. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. 2009 Oct 8; 461(7265):747–53. https://doi.org/10.1038/nature08494 PMID: 19812666
22. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet Med. 2002; 2:45–61.
23. Naj AC, Jun G, Beecham GW, Wang LS, Vardarajan BN, Buros J, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's. Nature Genetics. 2011 May; 43(5):436–41. https://doi.org/10.1038/ng.801 PMID: 21460841
24. Richie K, Carriere I, Richi CW, Berr C, Artero S, Ancelin ML. Designing prevention programs to reduce incidence of dementia: prospective cohort study of modifiable risk factors. British Medical Journal. 2010 Aug 5; 341:c3885. https://doi.org/10.1136/bmj.c3885 PMID: 20688841

25. Van Duijn CM, Hofman A. Relation between nicotine intake and Alzheimer's disease. British Medical Journal. 1991; 22:1491–1494.

26. Lee, C., Alekseenko, A., Brown, T. (2009) Exploring the future of bioinformatics data sharing and mining with Pygr and Worldbase, Proceedings of the 8th Python in Science Conference (SciPy 2009)

27. Hauck WW, Neuhaus JM, Kalbfleisch JD, Anderson S (1991) A consequence of omitted covariates when estimating odds ratios. J Clin Epidemiol; 44(1):77–81 PMID: 1986061

28. Hoffman A, Sportelli V, Ziller M, Spengler D. Driver or Passenger: Epigenomes in Alzheimer's disease. Epigenomes. 2017; 1(1) 5; https://doi.org/10.3390/epigenomes1010005