HHS Public Access

Author manuscript

Genet Epidemiol. Author manuscript; available in PMC 2019 September 01.

Published in final edited form as:

Genet Epidemiol. 2018 September ; 42(6): 551–558. doi:10.1002/gepi.22135.

GENE-ENVIRONMENT INTERACTIONS IN CASE-CONTROL STUDIES WITH SILENT DISEASE

Iryna Lobach[1,*], **Joshua Sampson**[2], **Siarhei Lobach**[3], and **Li Zhang**[4,1]

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

[2]National Cancer Institute, National Institutes of Health, Bethesda MD, USA

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

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

Abstract

Genome-Wide Association Studies (GWAS) often measure Gene-Environment (GxE) interactions. We consider the problem of accurately estimating a GxE interaction in a case/control GWAS when a subset of the controls have silent, or undiagnosed, disease and the frequency of the silent disease varies by the environmental variable. We show that using case/control status without accounting for misdiagnosis can lead to biased estimates of the GxE interaction. We further propose a pseudo-likelihood approach to remove the bias and accurately estimate how the relationship between the genetic variant and the *true* disease status varies by the environmental variable. We demonstrate our method in extensive simulations and apply our method to a GWAS of prostate cancer.

INTRODUCTION

We are interested in studying gene-environment interactions (GxE) in case-control Genome-Wide Association Studies (GWAS) where a substantial proportion of "controls" are actually undiagnosed cases. We note there are numerous diseases that go undiagnosed in a large segment of the population. For example, Atrial Fibrillation is undiagnosed in 5–17% of the population above the age of 75 (Panisello-Tafalla et al. 2015), non-alcoholic fatty liver disease is undiagnosed in 14–30% of the adult population (El-Kader et al., 2015), and acute coronary thrombosis is undiagnosed in >10% of individuals at the time of death (Anderson et al, 1989). These frequencies often vary by the environment, e.g. age, sex, race/ethnicity. Our specific motivating example is a large GWAS of prostate cancer. At autopsy, approximately 29%, 36%, and 47% of "healthy" men aged 60–69, 70–79 and 80+ years have undiagnosed prostate cancer, with the exact frequencies varying by race and ethnicity (Jahn et al, 2015).

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

There is already an extensive literature (Carroll et al, 2006) discussing how the estimates of the main effect of the genetic variant will be biased in the presence of undiagnosed controls. Here, we extend the literature by showing how the estimates of the GxE interaction can be biased when there is a relationship between the environmental variable and the rate of misdiagnosis. In our motivating example of Prostate Cancer, the environmental variable is age and the rate of misdiagnosis is known to increase with age. The result is that the effect of the gene would appear to vary by age, even when, in truth, there is no such interaction. After demonstrating the potential for bias in the GxE interaction, we propose a new method that uses external knowledge about the rates of misdiagnosis to accurately estimate the GxE interaction.

Our proposed method is based on the method of Chatterjee and Carroll (Chatterjee and Carroll, 2005). Initially (Thomas, 2010), GxE in case-control GWAS were analyzed using logistic regression, with the data treated as if it were collected prospectively (Prentice and Pyke, 1979). However, Chatterjee and Carroll showed that when the data is collected retrospectively and the gene and environmental variables are independent, there are more efficient methods for estimating the GxE interaction. We adapt these fully efficient methods to the scenario where the disease is undiagnosed in a subset of controls.

Our paper proceeds as follows. First, in the Material and Methods section, we describe our notation, the proposed pseudo-likeilhood approach and its properties. Next, in the Simulation Experiments section, we compare our proposed approach with standard approaches that ignore misdiagnosis. Then, we apply our approach to a Prostate Cancer GWAS (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000207.v1.p1, Yeager et al, 2007). Finally, we conclude our paper with a brief Discussion section.

MATERIALS AND METHODS

Notation and Estimation in Pseudo-likelihood

For individual i, let G_i be the genotype, X_i be the environmental variable potentially interacting with the genotype, and Z_i be a vector of other environmental variables. We will assume that the genotype is independent of all environmental variables and the genotype follows Hardy-Weinberg Equilibrium: $G \sim Q(g, \theta)$. Let $D_i = \{0, 1\}$ be a binary indicator of the *true*, and unobserved, disease status and let $D_i^{CL} = \{0, 1\}$ be a binary indicator of clinically diagnosed disease status. In the overall population, let $\pi_0 = pr(D^{CL} = 0)$ and $\pi_1 = pr(D^{CL} = 1)$ and in our study population let n_0 be the number of controls (i.e. $D^{CL} = 0$), n_1 be the number of cases (i.e. $D^{CL} = 1$), and $n = n_0 + n_1$.

We make the following assumptions. We assume that individuals with a clinical diagnosis have the true disease, i.e. $\text{pr}(D = 1 \mid D^{CL} = 1) = 1$, and that a substantial proportion of "controls" also have the true disease and that this proportion can vary by environmental

factors: $pr(D = 1 | D^{CL} = 0, X) = \tau(X) \gg 0$. We next assume that the probability of the *true* disease follows a logistic model

$$pr\ (D = 1 \Big| G = g, X = x, Z = z) \qquad 1$$
$$= \frac{\exp\{\beta_0 + \beta_X \times x + \beta_Z \times z + \beta_G \times g + \beta_{G \times X} \times g \times x\}}{1 + \exp\{\beta_0 + \beta_X \times x + \beta_Z \times z + \beta_G \times g + \beta_{G \times X} \times g \times x\}},$$

but note that our approach can be easily extended to other models, including those with multiple disease states.

The observed data are collected using retrospective sampling design were the genetic and environmental variables are measured after the disease status is ascertained. Instead, however, we imagine that individuals were selected into the study using the following Bernoulli scheme. Let δ be the imaginary indicator of whether an individual is selected into the case/control study with $pr\ (\delta = 1 | D^{cL} = d^{cl}) \propto n_{d^{cl}} / \pi_{d^{cl}}$. Let

$\kappa_{d^{cl}, d} = \beta_{0d} + \log\left(n_{d^{cl}} / \pi_{d^{cl}}\right), \gamma_{d^{cl} | d}(X) = pr\ \left(D^{cL} = d^{cl} | D = d, X\right),$

$\Omega = \left(\kappa_{d^{cl}, d}, \beta_0, \beta_X, \beta_z, \beta_G, \beta_{G \times X}, \theta\right)$ and

$$s(d, d^{cl}, g, x, z; \Omega) = \frac{\exp[I(d = 0) \times \left\{k_{d^{cl}, d} + \beta_X \times x + \beta_Z \times z + \beta_G \times g + \beta_{G \times X} \times g \times x\right\}]}{1 + \exp\left\{\beta_0 + \beta_X \times x + \beta_Z \times z + \beta_G \times g + \beta_{G \times X} \times g \times x\right\}} \times Q(g; \theta)$$

We now have the pseudo-likelihood constructed based on the probability of $\left[D^{cL}, G | X, Z, \delta = 1\right]$ in the following form

$$\prod_{i=1}^{N} L\left(d_i^{cl}, g_i, x_i, z_i; \Omega\right), \qquad 2$$

where

$$L\left(d^{cl}, g, x, z; \Omega\right) = \frac{s(0, 0, g, x, z; \Omega) + \gamma_{d^{cl} | 1}(x) \times S\left(1, d^{cl}, g, x, z; \Omega\right)}{\Sigma g^* \left\{S(0, 0, g^*, x, z; \Omega) + \sum_{d^{cl}_*} \gamma_{d^{cl}_* | 1}(x) \times \mathrm{S}\left(1, d^{cl_*}, g^*, x; \Omega\right)\right\}}.$$

Interestingly, the intercept parameter $k_{d^{cl}, d}$ is now a function of the probability of the clinical diagnosis in the population. Hence estimation can be improved by entering a reliable estimate or a bound on the probability of the clinical disease in the population that is often

available in epidemiologic studies. Furthermore, conditioning on $\{X, Z\}$ makes it possible to avoid specification of their distribution.

The use of pseudo-likelihood defined using (2) needs to be justified. Arguments provided in Appendix demonstrate that the parameter estimates obtained by maximizing this pseudo-likelihood (2) are consistent and that under suitable regularity conditions the parameters have the asymptotic variance-covariance matrix described below.

Define $\Psi(d^{cl}, g, x, z; \Omega)$ to be the derivative of $\log\left\{L(d^{cl}, g, x, z; \Omega)\right\}$ with respect to Ω and

$$\mathscr{L}_N(\Omega) = \sum_{i=1}^{N} \Psi\left(D_i^{cL}, G_i, X_i, Z_i; \Omega\right);$$

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

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

where all expectations are taken with respect to the actual retrospective sampling scheme. Then

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

SIMULATION EXPERIMENT

We compare three procedures for estimating the parameters in equation (1). The first is the usual logistic regression model (uLR) that uses clinical diagnosis as a surrogate for the *true* diagnosis. The second is the pseudolikelihood approach (pMLE) proposed by Chatterjee and Carroll (2006) where, again, the clinical diagnosis is used. The third is our approach (pMLE-DX) which accounts for the frequencies of silent disease by maximizing equation (2).

We simulate data assuming that the relationship between the *true* disease status and the combination of gene (G) age (X) and family history (Z) can be described by equation (1). In all simulations, we let G be a Bernoulli variable with probability 0.1. We let X be an unordered categorical variable that takes values 0, 0.8, 1 with probabilities 0.488, 0.165, and 0.347. The values of X are chosen to reflect main effect of age, and frequencies of categories of X are defined to reflect prostate cancer rates in the general population. We let Z be a Bernoulli variable with probability 0.07. We vary the coefficients, sample size, and frequency of undiagnosed disease in the four settings. For each simulation setting, we simulate 500 datasets and then look the bias and RMSE of the parameters using each of the three methods.

Setting 1.

We first simulate datasets assuming there is no GxE interaction and $\text{pr}(D = 1 \mid D^{CL} = 0, X)$=0.29 for X = 0, 0.36 for X = 1, and 0.44 for X=0.8. Furthermore, we let $\beta_0 = -1.035$, β_X=1, β_Z=2.5, $\beta_{G\times X}$=0 and vary β_G from log(1.05) to log(3.5). Table 1 describes performance of the three methods in studies where $n_0 = n_1 = 3000$; while Supplementary Table 1 describes the performance when in studies where $n_0 = n_1 \in \{1000, 5000\}$ The estimates of both β_G and $\beta_{G\times X}$ are biased when using uLR and pMLE, while the estimates for these parameters are effectively unbiased when using pMLE-DX. For example, when $n_0 = n_1 = 3000$ and β_G = log(2.5)=0.92 and $\beta_{G\times X} = 0$, the uLR and pMLE bias in $\widehat{\beta_G}$ is −0.34 and in $\widehat{\beta_{GxX}}$ the bias is −0.10; while the bias is only 0.009 and 0.007 when using pMLE-DX.

Setting 2:

We simulate datasets with $\beta_{G\times X} = 0.5$ and pr(D=1\D$^{\text{CL}} = 0.29$ for X=0, 0.44 forX=0.8, and 0.36 forX=1. Specifically, we let $\beta_0 = -1.035$, $\beta_G = -0.311$, $\beta_X = 1$, $\beta_Z = 2.5$. These parameters result in a prevalence of disease of 30%, 47.5%, 52.2% in $X = 0,\ 0.8,$ and 1. Furthermore, for setting 2a, we let the relationship between the *true* and clinically diagnosed disease statuses be defined by $\text{pr}(D = 1\backslash D^{CL} = 0, X)$=0.10 for X=0, 0.30 for X=0.8 and 0.5 for X=1 For setting 2b, we let the same relationship be defined by $\text{pr}(D = 1\backslash D^{CL} = 0, X)$=0.10 for X=0 , 0.30, for X=0.8 , and 0.5 for X=1 so that there is greater variability in the rate of misdiagnosis with lower frequency in the first category. Tables 2 (setting 2a) **and** 3 (setting 2b) describe of performance of the three methods in studies where $n_0 = n_1 = 3000$ while Supplementary Tables 2 (setting 2a) **and** 3 (setting 2b) describe the performance when in studies with $n_0 = n_1 \in \{1000,\ 5000\}$ As expected, in both settings 2a and 2b, pMLE-DX produces nearly unbiased estimates of all parameters. In contrast, both uLR and pMLE produce estimates of β_0, β_x, β_z, $\beta_{G\times X}$ that are biased. Note, these biases persist in the larger case/control studies with $n_0 = n_1 \in \{1000,\ 5000\}$ (Supplementary Table 3, setting 2b). The bias in standard approaches (uLR and pMLE) was larger when the frequency of silent disease was higher (setting 2a vs. 2b). For example, in setting 2a, the bias of $\widehat{\beta_X}$ is 1.9, the bias of $\widehat{\beta_Z}$ is −1.5, and the bias of $\widehat{\beta_{GxX}}$ is −0.11, while in setting 2b the corresponding biases are −0.74, −0.31 and 0.05.

Setting 3:

We wanted to assess the effect of misspecifying $\tau(X)$, the relationship between the *true* and clinically diagnosed disease statuses. We therefore simulated data as described in setting 1, but misspecified the $\tau(X)$ = 0.29 for all three age groups. As shown in Supplementary Table 4, pMLE-DX no longer resulted in unbiased estimates, but the bias was significantly lower than when using either of the other two methods.

Setting 4. β_Z **and** β_X:

To better understand the nature of the bias noted in estimates of β_Z and β_X, we conducted a study with these parameters varying from 0 to 2.5 in steps of 0.5. The other coefficients are set to be $\beta_0 = -1.05, \quad \beta_G = \log(3), \quad \beta_{G\times X} = \log(2)$. The relationship between the clinical and *true* disease status was $\text{pr}\left(D = 1|D^{CL} = 0, X\right) = 0.10$ for $X = 0$, 0.20 for $X = 0.8$ and 0.30 for $X = 1$. Supplementary Figures 1 **and** 2 present the probabilities of the *true* disease status and clinically diagnosed disease status across the different values of β_Z and β_X. When $\beta_X = 0$, the probabilities of the *true* disease status vary only slightly across X categories, while the probabilities of the clinically diagnosed disease status vary substantially across X categories because the frequency of silent disease varies by X. Figures 1 A and B present the bias and RMSE of $\widehat{\beta_Z}$ obtained by uLR with color-coded values of β_Z and values of β_X on x-axis. As the *true* value of β_Z increases, the bias in its estimates increases as well. Figure 1 C and D present $\widehat{\beta_X}$ obtained by uLR with color-coded values of β_X and values of β_Z on x-axis. The Bias in $\widehat{\beta_X}$ decreases when the true value increases.

Setting 5. $\beta_X = 0$ and $\beta_{G\times X} = 0$**:**

To better elucidate the underlying nature of biases noted in estimates of β_X and β_Z obtained in uLR, we performed a simulation study varying the relationship between *true* and clinical disease statuses. For clarity, X is now simulated as Bernoulli with frequency 0.52. Both Z and G are now Bernoulli with frequencies 0.07 and 0.10, respectively. The risk coefficients are $\beta_0 = -1.05, \quad \beta_X = 0, \quad \beta_Z = 1, \quad \beta_{G\times X} = 0$. The study consists of 3,000 cases and 3,000 controls. First, we examined situations when the frequency of the silent disease does not vary by X, i.e.$\Delta = \text{pr}\left(D = 1|D^{CL} = 0, \text{X} = 0\right) - \text{pr}\left(D = 1|D^{CL} = 0, \text{X} = 1\right) = 0$ and when the difference is 0.05 and 0.10. We varied $\text{pr}\left(D = 1|D^{CL} = 0, \text{X} = 1\right)$ from 0.5 to 0.25 in steps of 0.05. Shown on Figure 2 A-B are biases and RMSE of estimates of β_Z across color-coded values of $\text{pr}\left(D = 1|D^{CL} = 0, \text{X} = 1\right)$ and across differences Δ on the x-axis. Similarly, shown on Figure 2 C-D are biases and RMSEs of estimates of β_X. When the clinical diagnosis is a surrogate of the *true* diagnosis (x-axis= Δ=0), then both $\widehat{\beta_X}$ and $\widehat{\beta_Z}$ are nearly unbiased. When the frequency of the silent disease varies more by age as Δ increases to 0.05 and 0.10, there is a notable increase in the bias and RMSE of $\widehat{\beta_X}$, while bias in $\widehat{\beta_Z}$ remains approximately the same. The bias and RMSE in estimates of $\widehat{\beta_X}$ increase with the proportion of undiagnosed controls. When frequencies of the silent disease increase, the biases in both $\widehat{\beta_Z}$ and $\widehat{\beta_X}$ increase with the bias in $\widehat{\beta_X}$ taking on a more rapid increase.

PROSTATE CANCER DATA ANALYSES

We performed GxE analyses for Prostate Cancer using data collected as part of the Prostate, Lung, Colon and Ovarian (PLCO) Screening trial (dbGAP: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000207.v1.p1, study accession phs000207.v1.p1, Yeager et al, 2007). The study included 965 cases and 1,035 controls of European ancestry with 550,000 genotyped SNPs. The number of cases in 50–59, 60–69, 70–79, and 80–89 year age groups were 111, 525, 326 and 3, respectively; the number of controls in same groups were 129, 598, 306 and 2. Furthermore, 11.3% of cases and 6.2% of controls had a family history of prostate cancer. In the following, we focused on the 81 SNPs that have well established associations documented in the National Human Genome Research Institute and the European Bioinformatics Institute (NHGRI-EBI) GWAS catalog (https://www.ebi.ac.uk/gwas/search?query=prostate) as of April 26, 2017. For each of the 81 SNPs, we assumed the relationship between the *true* disease status and the combination of SNP, family history, and age can be described by logistic regression and equation (3).

$$\begin{aligned}\text{logit}\{\text{pr}(D=1|Age,\quad FamHist,G)\} = \quad & \beta_0+\beta_{Age}\times Age+\beta_{FamHist}\times FamHist \\ & +\beta_{FamHistxAge}\times FamHist\times Age+\beta_G\times G+\beta_{GxAge}\times G\times Age\,.\end{aligned} \quad (3)$$

As in the simulations, we estimate the coefficients using uLR, pMLE and pMLE-DX where we assumed the relationship between clinical disease status and the *true* disease status is

$\text{pr}\left(D=1|D^C=0,\text{Age}\right)=0.22,\quad 0.30,\quad 0.35$and 0.46 for age groups of 50–59, 60–69, 70–79 and ≥ 80 years, respectively (Jahn et al, 2015). Statistical significance was assessed using permutation-based p-value <0.05.

The estimates of coefficients based on the three methods are quantitatively different. Table 4 and Supplementary Table 5A present the estimates of the risk coefficients for the 18 SNPs with p-value <0.05 for $\widehat{\beta}_{GxAge}$ in pMLE-DX. Values of all risk coefficient estimates are generally larger in the relationship to the *true* (pMLE-DX) than to the clinical disease status (uLR, pMLE). And the estimates are similar in uLR and pMLE. For example, consider rs103294 that is located on LILRA3, a key component in the regulation of inflammatory inhibition. This SNP was previously identified as being significantly associated with prostate cancer in a Chinese population (OR=1.28) (Xu et al, 2012). In our analyses of the PLCO dataset, rs103294 is estimated to be related to the observed clinical diagnosis with main effect OR=1.26 in uLR and pMLE and is associated with the latent *true* diagnosis with main OR= 1.39 in pMLE-DX. In addition, in pseudolikelihood analyses (pMLE-DX), the effect of the interaction between this SNP and age was found to be significant with p-value <0.05/81 and OR = 1.5. Estimates of the effect of age are OR=1.9, p=0.015 and OR=2.5, p=0.014 in uLR, pMLE and pMLE-DX, respectively. Estimates of the effect of family history are OR=1.4, p=0.001 and OR=1.1, p<0.001 in uLR and pMLE-DX, respectively. Supplementary Table 5B shows the estimates of the additional 16 SNPs whose $\hat{\beta}_G$ has p-value <0.05

(pMLE-DX) with similar tendency in the estimates, i.e. the risk coefficient estimates are generally larger in the relationship to the *true* disease status than the clinical disease status.

Four SNPs (rs339331 (6p22), rs1983891 (6p21), rs7501939 (17q12), rs6983267 (8q24),) have p-values <0.05/81 for $\hat{\beta}_G$ for both the clinical and *true* disease statuses. Two of these SNPs, rs339331 (6p22) and rs1983891 (6p21), were previously found to be associated with prostate cancer risk in an Asian population as well as European descent (Lindstrom et al, 2012) with OR=0.93, p=0.002 and OR=1.09, p=2.48×10^{-4}. In our analyses OR = 1.8 and 2.4 as a result of uLR and pMLE-DX. The third SNP, rs7501939 (17q12), was previously found to contribute to risk of early-onset prostate cancer (Levin et al, 2008 and references herein) with OR=1.19–1.44, p<0.008. In our analyses, OR was estimated to be 1.7 and 2.4 in uLR and pMLE-DX. The fourth SNP, rs6983267, is one of five SNPs used (with family history as a sixth factor) to cumulatively predict the overall risk of the diagnosis (Zheng, et. Al, 2008). On its own, the rs6983267(G;G) and (G;T) risk genotypes yield an odds ratio for developing prostate cancer of 1.37, p=3.4–10e-5) and were estimated to account for 22.2% of population attributable risk (Zheng et al, 2008). In our analyses the estimated OR are 1.7 and 2.2 in uLR and pMLE. While the estimates of uLR and pMLE are approximately the same, p-values as a result of pMLE are generally smaller. For example, as shown in Table 4, 3 SNPs have p-value for $\widehat{\beta}_{GxAge}$ that are <0.05/81 in pMLE, while for the other two approaches the p-value is >0.01.

DISCUSSION

We examined the potential bias in the estimates of the gene-environment interaction in the situation when a substantial portion of the population carries so-called silent disease that is not visible clinically. The bias arises from the relationship between *true* disease and clinical diagnosis varying across the environmental variable. We showed, both in simulation experiments and in the data analyses, that the potential bias can either over- or under-estimate the true interaction and that magnitude of this bias can be substantial. Moreover, this bias cannot be eliminated by increasing the sample size.

Others have also investigated the effect of disease misclassification in GWAS, albeit focusing on the bias occurring in the main effect. Their conclusions were similar, in that the bias can be in either direction and the magnitude of the bias can be substantial. For example, a recent study by Rekaya et al (2016) examined biases in the main estimates of the genetic factors when disease misclassification varies by the diagnosis. In our scenario, we also had disease misclassification varying by diagnosis (i.e. $\mathrm{pr}\left(D=1|D^{CL}=0,X\right)\backslash ne\mathrm{pr}\left(D=0|D^{CL}=1,X\right)$,, but the misclassification rate varied by an environmental variable.

The proposed analyses rely on knowing the estimates of silent disease in the population subgroups. These estimates are often available in epidemiologic studies or can be estimated in an internal reliability study. We found that when the frequency of the silent disease is either under– or over-estimated by less than 5%, then ignoring the presence of the silent disease still results in a bias that is larger than observed using our proposed approach. In

general, we advocate for a sensitivity analyses that varies the frequency of the silent disease and examines the resulting differences in the estimates.

We note that the potential errors in estimating effects can have downstream consequences. By underestimating the GxE interaction, we would also underestimate the heritability explained by SNPs. Therefore, the prior use of simple logistic regression to estimate interactions might contribute to the problem of missing heritability, or the difference between GWAS-based and family-based estimates of heritability (Manolino et al, 2009). On the other hand, the upward biases in these estimates might in part address the conclusion reached by Hirschhorn et al (2002) that only 1% of the association found are likely to be true.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Prostate cancer dataset was downloaded from the database of genotypes and phenotypes (https://dbgap.ncbi.nlm.nih.gov), study accession number phs000297.v1.p1.

https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000207.v1.p1.

We thank Ivan Belousov for help with the computations.

LITERATURE CITATIONS

Anderson RE, Hill RB, Key CR (1989) The sensitivity and specificity of clinical diagnostics during five decades: toward an understanding of necessary fallibility, JAMA, 261: 1610–1617.10.1001/jama.1989.0320110086029 [PubMed: 2645451]

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

Chatterjee N and Carroll RJ (2005) Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies, Biometrika, 92,2 399–418

El-Kader SMA and El-Den Ashmawy (2015) Non-alcoholic fatty liver disease: the diagnosis and management, World Journal of Hepatology, 7(6): 846–858 [PubMed: 25937862]

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. (2002) A comprehensive review of genetic association studies. Genet Med. 2:45–61

Jahn JL, Giovannucci EL, Stampfer MJ (2015) The high prevalence of undiagnosed prostate cancer at autopsy: implications for epidemiology and treatment of prostate cancer in the Prostate-specific Antigen-era. International Journal of Cancer, 137, 2795–2802 [PubMed: 25557753]

Levin AM,Machiela MJ,Zuhlke KA,Ray AM,Cooney KA,Douglas. (2012) Chromosome 17q12 variants contribute to risk of early-onset prostate cancer, Cancer Research, 15:68(16):6492–5

Lindström S,Schumacher FR,Campa D,Albanes D,Andriole G,Berndt SI…Kraft P. (2012) Replication of five prostate cancer loci identified in an Asian population--results from the NCI Breast and Prostate Cancer Cohort Consortium (BPC3). Cancer Epidemiol Biomarkers Prev, 21(1):212–6 [PubMed: 22056501]

Lobach I, Carroll RJ, Spinka C, Gail MH, Chatterjee N (2008) Haplotype-based regression analysis and inference of case-control studies with unphased genotypes and measurement error in environmental exposures, Biometrics, 64 673–684 [PubMed: 18047538]

Manchia M, Cullis J, Turecki G, Rouleau GA, Uher R, Alda M (2013) The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases, PLOS One, 8(10) 1–7

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, … Visscher PM(2009) Finding the missing heritability of complex diseases. Nature; 461:747–753. [PubMed: 19812666]

MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, …, Flicek P, Cunningham F, and Parkinson H. (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Research, 2017, Vol. 45 (Database issue): D896–D901

Panisello-Tafalla A, Clua-Espuny JLC, Gil-Guillen VF, Gonzalez-Henares A, Queralt-Tomas ML, Lopez-Pablo C, …Lopez MG (2015) Results from the registry of Atrial Fibrillation (AFABE): Gap between undiagnosed and registered atrial fibrillation in adults – ineffectiveness of oral anticoagulation treatment with VKA, Biomedical Research International, Vol 2015, 134756

Prentice KL and Pyke DA (1979) Logistic disease incidence models and case-control studies, Biometrika, Vol 66, 3, 403–411

Thomas D (2010) Gene-Environment-Wide Association Studies: emerging approaches, Nature Review Genetics, 11(4):259–272

Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, …, Thomas G. Genome-Wide Association Study of Prostate Cancer Identifies a Second Locus at 8q24. (2007) Nature Genetics, 39(5): 645–649 [PubMed: 17401363]

Xu J, Mo Z, Ye D, Wang M, Liu F, Jin G, …, Sun Y. (2012) Genome-wide association study in Chinese men identifies two new prostate cancer risk loci at 9q31.2 and 19q13.4, Nature Genetics, 44, 1231–1235 [PubMed: 23023329]

Zheng SL,Sun J,Wiklund F,Smith S,Stattin P,Li G,…,Grönberg H. (2008) Cumulative association of five genetic variants with prostate cancer, New England Journal of Medicine, 28:358 (9):910–9

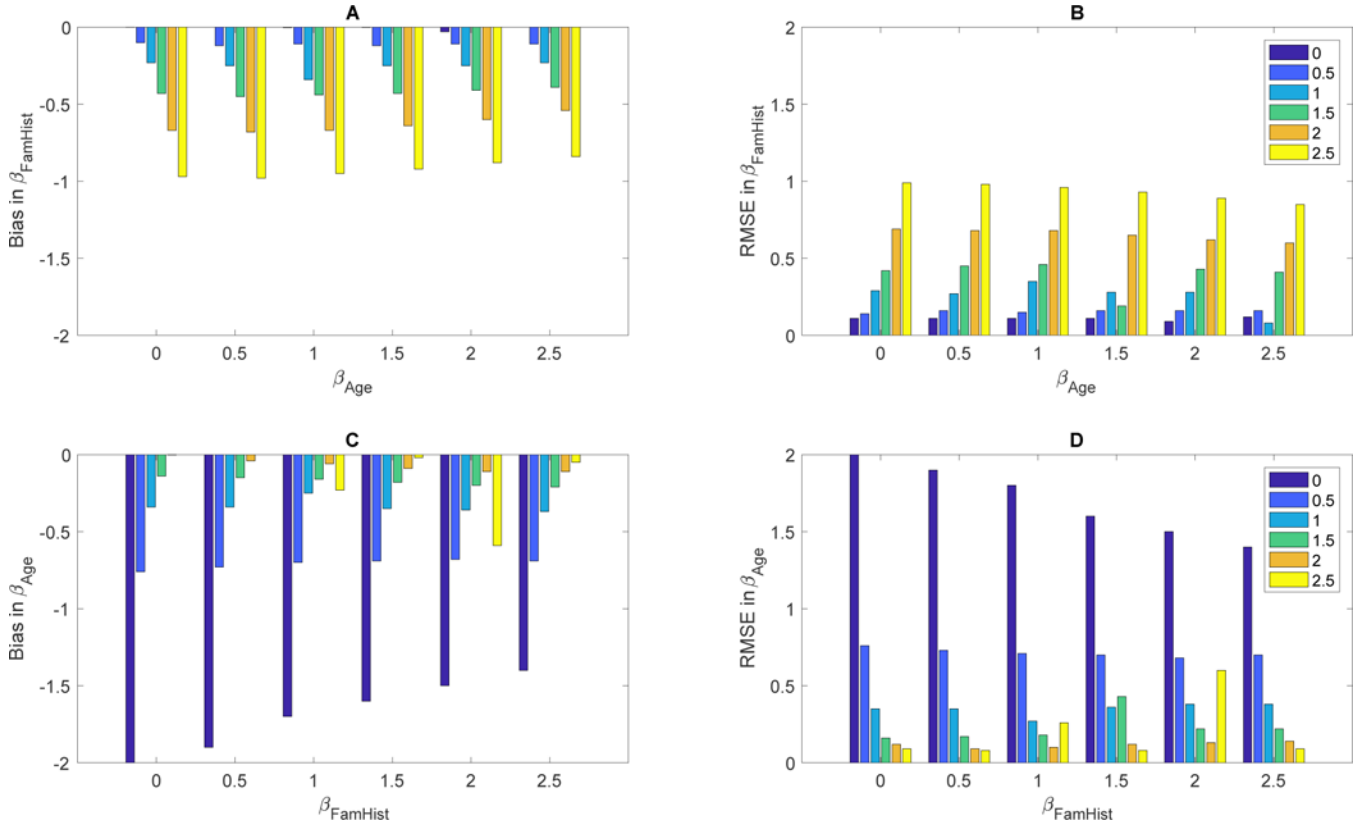

Figure 1:

The bias and Root Mean Squared Error (RMSE) in estimates of β_Z ($\beta_{FamHist}$) (**A-B**) and β_X $\left(\beta_{Age}\right)$ (**C-D**) obtained using the usual logistic regression with clinical diagnosis as the outcome. In **A-B,** the true values of β_X $\left(\beta_{Age}\right)$ are listed along the x-axis and the true values of β_Z $\left(\beta_{FamHist}\right)$ are indicated by color. On **C-D,** the true values of β_Z $\left(\beta_{FamHist}\right)$ are listed along the x-axis and the true values of β_X $\left(\beta_{Age}\right)$ are indicated by color. The parameters are set as follows: $\beta_0 = -1.05$, $\beta_G = \log(3)$, $\beta_{G\times X} = \log(2)$; the relationship between the clinical and *true* disease statuses is $\text{pr}\left(D = 1|D^{CL} = 0, \text{X}\right) = 0.10$ for $X = 0$, 0.20 for $X = 0.8$, and 0.30 for $X = 1$; both G and Z are Bernoulli with frequencies 0.10 and 0.07; variable X is multinomial with frequencies 0.488 for $X = 0$, 0.165 for $X = 0.8$ 0.30 for $X = 1$; n_0=n_1=3000. The frequencies of the latent *true* diagnosis and the observed clinical diagnosis are shown on Supplementary Figures 1 and 2.

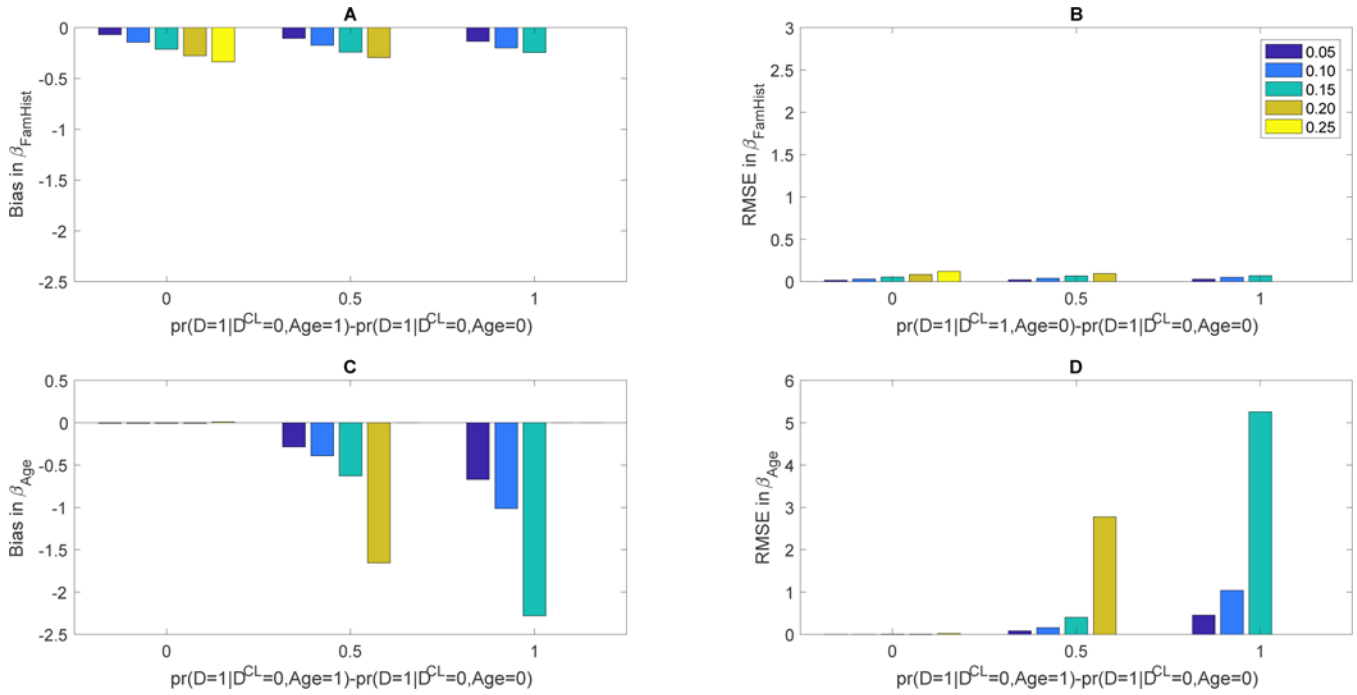

Figure 2:

The bias and Root Mean Squared Error (RMSE) in β_Z $(\beta_{FamHist})$ (**A-B**) and β_X (β_{Age}) (**C-D**) obtained in uLR across 500 simulated datasets with 3,000 cases and 3,000 controls when the clinical diagnosis is used in place of the *true* diagnosis. Variables G, X (Age), and Z $(FamHist)$ are Bernoulli with frequencies 0.10, 0.52, 0.07, respectively. The risk coefficients are $\beta_0 = -1.05$, $\beta_X = 0$, $\beta_Z = 1$, $\beta_{G\times X} = 0$. Values of the difference $\text{pr}\left(D = 1|D^{CL} = 0, X = 1\right) - \text{pr}\left(D = 1|D^{CL} = 0, X = 0\right)$ are listed along the x-axis and values of $\text{pr}\left(D = 1|D^{CL} = 0, X = 1\right)$ are indicated by color. Shown are only parameter values that keep probability of the clinical diagnosis and probability of *true* diagnosis within 0 to 1 range.

Table 1:

The Bias and Root Mean Squared Error (RMSE) in estimates of from $\beta_{G\times X}$ and β_G 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 500 datasets with n_0 = 3000 controls and n_1 = 3000 cases; the variable X takes on values 0, 0.8, 1 with probabilities 0.488, 0.165 and 0.347; the relationship between the clinical and *true* diagnoses varies by X, i.e.$\text{pr}\left(D=1|D^{CL}=0,\text{X}\right)=0.29$ for $X=0$; 0.474 for $X=0.8$, and 0.357 for $X=1$; the risk of disease follows equation 2 with $\beta_0=-1.035$, $\beta_X=1.00$, $\beta_Z=2.5$, $\beta_{G\times X}=0$; $\text{pr}(G=1)=0.10$.

$\beta_{G\times X}=0$	Estimates of $\beta_{G\times X}$						Estimates of β_G					
	Usual Logistic Regression		pMLE		pMLE-DX		Usual Logistic Regression		pMLE		pMLE-DX	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
$\beta_G=$	***n_0=n_1=3,000***											
log(1.1)	0.017	**0.86**	0.013	**0.25**	0.029	**0.37**	−0.04	**0.86**	−0.04	**0.21**	−0.006	**0.29**
log(1.2)	0.016	**0.83**	0.023	**0.24**	0.037	**0.35**	−0.09	**0.83**	−0.09	**0.23**	−0.03	**0.29**
log(1.3)	−0.016	**0.81**	−0.02	**0.24**	0.002	**0.36**	−0.11	**0.81**	−0.11	**0.22**	−0.02	**0.30**
log(1.4)	−0.02	**0.77**	−0.02	**0.22**	0.017	**0.35**	−0.12	**0.77**	−0.12	**0.22**	−0.003	**0.27**
log(1.5)	−0.04	**0.73**	−0.04	**0.22**	0.003	**0.35**	−0.14	**0.73**	−0.14	**0.23**	−0.004	**0.28**
log(2.0)	−0.07	**0.60**	−0.08	**0.17**	0.017	**0.37**	−0.24	**0.60**	−0.25	**0.30**	0.018	**0.27**
log(2.5)	−0.10	**0.50**	−0.11	**0.19**	0.009	**0.41**	−0.34	**0.50**	−0.35	**0.38**	0.007	**0.28**
log(3.0)	−0.11	**0.42**	−0.12	**0.19**	0.04	**0.42**	−0.45	**0.42**	−0.45	**0.47**	−0.023	**0.26**
log(3.5)	−0.15	**0.37**	−0.15	**0.21**	0.03	**0.47**	−0.49	**0.37**	−0.49	**0.51**	0.031	**0.29**

Table 2:

The Bias and Root Mean Squared Error (RMSE) in estimates of $\beta_{G\times X}$ and β_G 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 500 datasets with n_0 = 3000 controls and n_1 = 3000 cases; the variable X takes on values 0, 0.8, 1 with probabilities 0.488, 0.165 and 0.347; the relationship between the clinical and *true* diagnoses varies by X, i.e. $\text{pr}\left(D=1|D^{CL}=0,\text{X}\right)=0.29$for $X=0$; 0.474 for $X=0.8$, and 0.357 for $X=1$; the risk of disease follows equation (2) with $\beta_0=-1.035, \beta_G=\log(1.35), \beta_X=1.0, \beta_Z=2.5, \beta_{G\times X}=0;\ \ pr\ (G=1)=0.10$.Variable Z is Bernoulli with frequency 0.07.

Parameters	True value	Clinical diagnosis is used in place of *true* disease				Frequency of silent disease in controls is estimated in an external study	
		Usual logistic regression		Pseudo-likelihood (pMLE)		Pseudo-likelihood (pMLE-DX)	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
$n_0=3,000$and $n_1=3,000$							
β_0	−1.035	−1.08	1.10	−3.2	3.24	−0.001	0.05
β_G	0.311	0.05	0.30	0.06	0.28	−0.008	0.35
β_X	1.0	1.87	1.88	1.87	1.89	0.002	0.05
β_Z	2.5	−1.52	1.52	−1.51	1.52	0.17	0.80
$\beta_{G\times X}$	0.50	−0.11	0.33	−0.13	0.33	0.008	0.41
Pr(G=1)	0.10			0.02	0.02	0.02	0.03

Table 3:

The Bias and Root Mean Squared Error (RMSE) in estimates of $\beta_{G\times X}$ and β_G from simulations using the usual logistic regression with clinical diagnosis as the outcome (uLR), the pseudolikelihood approach (pMLE), and our newly proposed pseudo-likelihood approach that accounts for misdiagnosis (pMLE-DX). For these simulations, the study included 500 datasets with n_0 = 3000 controls and n_1 = 3000 cases; the variable X takes on values 0, 0.8, 1 with probabilities 0.488, 0.165 and 0.347; the relationship between the clinical and *true* diagnoses varies by X , i.e. $\text{pr}\left(D = 1|D^{CL} = 0, \text{X}\right) = 0.10$ for $X = 0$; 0.30 for $X = 0.8$, and 0.44 for $X = 1$; the risk of disease follows equation (2) with $\beta_0 = -1.035, \beta_G = \log(1.35), \beta_X = 1.0, \beta_Z = 2.5, \beta_{G\times X} = 0;\ \ pr\ (G = 1) = 0.10$. Variable z is Bernoulli with frequency 0.07.

Parameters	True value	Clinical diagnosis is used in place of *true* disease				Frequency of silent disease in controls is estimated in an external study	
		Usual logistic regression		Pseudo-likelihood (pMLE)		Pseudo-likelihood (pMLE-DX)	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
$n_0 = 3,000$ and $n_1 = 3,000$							
β_0	−1.035	0.84	0.84	−0.50	0.50	0.001	0.02
β_G	0.311	−0.04	0.12	−0.04	0.12	−0.005	0.14
β_X	1.0	−0.74	0.35	−0.75	0.32	−0.0003	0.03
β_Z	2.5	−0.31	0.75	−0.31	0.75	0.003	0.15
$\beta_{G\times X}$	0.50	0.05	0.19	0.05	0.19	0.005	0.24
Pr(G=1)	0.10			0.02	0.02	0.02	0.02

Table 4:

The estimates of effects and their p-values of individual SNPs (β_G), age (β_{Age}), family history ($\beta_{FamHist}$), and the interactions between family history and age ($\beta_{FamHist}$) and between SNP and age ($\beta_{FamHist}$) in the GWAS of prostate cancer. Estimates were obtained by 1) the usual logistic regression (uLR) model with clinical disease status as the outcome, 2) the usual pseudo-likelihood model with clinical disease as the outcome (pMLE), and 3) the newly proposed pseudo-likelihood model which accounts for misdiagnosis (pMLE-DX). The relationship between the clinical and the *true* disease statuses is defined as follows: $\text{pr}\left(D = 1 | D^C = 0, \text{Age}\right) = 0.22$for 50–59, 0.30 for 60–69, 0.35 for 70–79 and 0.46 for ≥ 80. Included are SNPs with permutation-based p-value for GxAge interaction <0.05 in pMLE-DX.

SNP	Model	β_G	β_{Age}	$\beta_{FamHist}$	$\beta_{FamHistxAge}$	β_{GxAge}
rs6983267	uLR pMLE pMLE-DX	**0.53**, p=0.000 0.53, p=0.000 **0.78**, p=0.000	**0.29**, p=0.004 **0.29**, p=0.000 1.0, p=0.13	**0.60**, p=0.0008 **0.60**, p=0.000 **0.91**, p=0.002	0.09, p=0.40 0.09, p=0.08 0.48, p=0.20	**−0.48**, p=0.02 **−0.48**, p=0.000 **−0.70**, p=0.02
rs11740657	uLR pMLE pMLE-DX	**−0.23**, p=0.03 **−0.23**, p=0.006 **−0.30**, p=0.05	0.06, p=0.29 **0.06**, p=0.000 **0.69**, p=0.02	**0.65**, p=0.001 **0.65**, p=0.000 **1.03**, p=0.000	0.01, p=0.47 0.01, p=0.18 0.32, p=0.29	**0.48**, p=0.01 **0.48**, p=0.000 **0.72**, p=0.02
rs8008270	uLR pMLE pMLE-DX	−0.22, p=0.05 **−0.22**, = 0.000 **−0.32**, p=0.03	0.08, p=0.25 **0.08**, p=0.000 **0.69**, p=0.01	**0.64**, p=0.0006 **0.64**, p=0.000 **1.1**, p=0.006	0.05, p=0.44 **0.05**, p=0.15 0.38, p=0.27	**0.57**, p=0.009 **0.57**, p=0.000 **0.97**, p=0.01