American Journal of Epidemiology
Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health 2021.
This work is written by (a) US Government employee(s) and is in the public domain in the US.

Vol. 190, No. 9
https://doi.org/10.1093/aje/kwab124
Advance Access publication:
May 2, 2021

Practice of Epidemiology

A Robust Test for Additive Gene-Environment Interaction Under the Trend Effect of Genotype Using an Empirical Bayes-Type Shrinkage Estimator

Nilotpal Sanyal, Valerio Napolioni, Matthieu de Rochemonteix, Michaël E. Belloy, Neil E. Caporaso, Maria Teresa Landi, Michael D. Greicius, Nilanjan Chatterjee, and Summer S. Han*

* Correspondence to Dr. Summer Han, Quantitative Sciences Unit, Department of Medicine, Stanford University School of Medicine, 1701 Page Mill Road, Palo Alto, CA 94304 (e-mail: summerh@stanford.edu).

Initially submitted July 1, 2020; accepted for publication April 21, 2021.

Evaluating gene by environment (G × E) interaction under an additive risk model (i.e., additive interaction) has gained wider attention. Recently, statistical tests have been proposed for detecting additive interaction, utilizing an assumption on gene-environment (G-E) independence to boost power, that do not rely on restrictive genetic models such as dominant or recessive models. However, a major limitation of these methods is a sharp increase in type I error when this assumption is violated. Our goal was to develop a robust test for additive G × E interaction under the trend effect of genotype, applying an empirical Bayes-type shrinkage estimator of the relative excess risk due to interaction. The proposed method uses a set of constraints to impose the trend effect of genotype and builds an estimator that data-adaptively shrinks an estimator of relative excess risk due to interaction obtained under a general model for G-E dependence using a retrospective likelihood framework. Numerical study under varying levels of departures from G-E independence shows that the proposed method is robust against the violation of the independence assumption while providing an adequate balance between bias and efficiency compared with existing methods. We applied the proposed method to the genetic data of Alzheimer disease and lung cancer.

additive risk model; Alzheimer disease; case-control design; empirical Bayes; gene–*APOE*-ε4 interaction; gene-environment interaction; gene-smoking interaction; GWAS

Abbreviations: APOE, apolipoprotein E; CML, constrained maximum likelihood; G-E, gene-environment; GWAS, genome-wide association study; LOAD, late-onset Alzheimer's disease; MAF, minor allele frequency; MOR, marginal odds ratio; RERI, relative excess risk due to interaction; SNP, single-nucleotide polymorphism; UML, unconstrained maximum likelihood.

Understanding the interaction between genes and environmental factors is important; genes do not function in isolation but rather in complex pathways influenced by environmental factors. There has been a long-standing controversy regarding the definition of interaction and the selection of proper scales for measuring the presence of interactions (1–3). Additive interaction measures the departure from a risk model that assumes that gene and environment act additively on the risk of the disease itself. Despite the popularity of the interaction tests under multiplicative risk models (that assume the multiplicative effects of gene and environment) via logistic regression, additive interaction has gained wider attention recently (4–10). This is partly due to its direct relevance for public health decision-making—such as whether it is beneficial to target individuals for intervention for an exposure based on genetic susceptibility—for which assessing absolute risk differences versus relative risks is more relevant (1, 3, 11, 12). In addition, additive interactions might shed light on biologic mechanisms when motivated by specific biologic hypotheses (12), although elucidating biological interactions using an additive or multiplicative model cannot be easily done mechanically (12, 13). Further, utilizing an assumption of the independence between gene and environment in the underlying population has been shown to yield a more precise estimate of G × E (gene-environment) interaction (5, 7, 8, 10, 14–17) for evaluating both additive (5, 7, 8, 10) and multiplicative interactions (14–17) in case-control studies.

Several methods have been proposed for evaluating additive G × E or gene-gene (G × G) interaction in the recent literature (5, 7, 9, 18–21). Specifically, Han et al. (5) developed a likelihood ratio test for additive interaction using a set of constraints for the joint effect of gene and environment under an additive risk model; this method utilizes the G-E independence assumption based on a retrospective likelihood framework (17) to boost power. However, this approach is based on strong assumptions of the underlying genetic models such as dominant or recessive effects that are known to be less robust when the true genetic model is unknown. While the use of a general genetic model with more than 2 categories of genotype has also been proposed for testing additive interaction, this approach has been shown to reduce power because of increased degrees of freedom (5). A recent study (9) has extended the likelihood ratio test from Han et al. by incorporating the trend effect (i.e., the linear effect of the genotype) of genotype, which is more robust across different underlying genetic models. However, the main limitation of this approach is the reliance on the strong assumption of G-E or G-G independence, the violation of which can lead to large type I error (22). In particular, the assumption of G-G independence is often inappropriate when interaction tests are applied to a group of single-nucleotide polymorphisms (SNPs) that are in high linkage disequilibrium. While more robust estimators of interaction have previously been proposed for multiplicative models (16) or for dominant/recessive effects of genotype under additive risk models (7) to address this issue, no such effort has been made for evaluating additive interaction under the trend effect of genotype.

In this work, we have proposed a robust statistical test for additive G × E or G × G interaction under the trend effect of genotype, applying an empirical Bayes–type shrinkage estimator to relax the strong independence assumption. We conducted a simulation study to evaluate the overall performance of the proposed method by estimating type I error, bias, and mean squared error under varying levels of departures from G-E independence. We applied the proposed method to analyze gene–apolipoprotein E (*APOE*)-ε4 interaction in data on late-onset Alzheimer disease (LOAD) and to examine gene-smoking interaction in lung cancer data. We have implemented our method in the R (R Foundation for Statistical Computing, Vienna, Austria) package CGEN (https://bioconductor.org/packages/release/bioc/html/CGEN.html).

METHODS

Additive risk model, additive interaction, and null hypothesis

Suppose, for subject i, G_i is a genetic risk factor (G) denoting the number of a minor allele of a bi-allelic SNP ($G_i = 0$, 1, or 2); E_i is a binary variable for an environmental risk factor (E), with $E_i = 0$ or 1 based on the presence or absence of the factor; $\mathbf{x}_i$ is the matrix of the covariates that can be included in a model; and D_i is a binary variable with $D_i = 0$ or 1 denoting the presence or absence of a disease, respectively. Under a general genetic model, without assuming the trend effect of genotype for now, the genetic factor variable G_i can be treated as a categorical variable and coded using a set of 2 dummy variables, G_{1i} and G_{2i}, that indicate whether the subject i has 1 copy ($G_{1i} = 1$) or 2 copies ($G_{2i} = 1$) of the minor allele. An additive risk model assumes that G_i and E_i act additively on the disease risk itself, hence without any link function (i.e., identity link function):

$$P(D_i = 1 \mid G_i, E_i, \mathbf{x}_i) = b_0 + b_{G_1} G_{1i} + b_{G_2} G_{2i} + b_E E_i + \mathbf{x}_i^T \mathbf{b}_\mathbf{x}.$$

A departure from the above additive risk model (i.e., additive interaction) can be explicitly modeled using the following extended model:

$$\begin{aligned} P\left(D_i = 1 \mid G_i, E_i, \mathbf{x}_i\right) &= b_0 + b_{G_1} G_{1i} + b_{G_2} G_{2i} + b_E E_i \\ &\quad + b_{G_1 E} G_{1i} E_i + b_{G_2 E} G_{2i} E_i + \mathbf{x}_i^T \mathbf{b}_\mathbf{x}, \end{aligned} \tag{1}$$

and the interaction can be tested using the null hypothesis of $H_0 : b_{G_1 E} = b_{G_2 E} = 0$. Although the parameters of the above model can be estimated using ordinary least squares, it can lead to predicted probabilities greater than 1 or less than 0, and it is difficult to account for ascertainment (e.g., case-control sampling). The usual solution is to reformulate the testing problem using the saturated logit model under a rare-disease assumption (5, 9). That is, suppose, $R_{ge} = P(D = 1 \mid G = g, E = e)$ denotes the disease risk for $g = 0, 1, 2$ and $e = 0, 1$ for any subject. Table 1 shows the disease risk for each combination of gene and environment, (g, e). The null hypothesis of no additive interaction, $H_0 : b_{G_1 E} = b_{G_2 E} = 0$, is equivalent to $H_0 : R_{10} - R_{00} = R_{11} - R_{01}$ and $R_{20} - R_{00} = R_{21} - R_{01}$, which implies that the change in disease risk for $G = 1$ versus $G = 0$ is the same across all levels of environment, and a similar relation holds for the risk change for $G = 2$ versus $G = 1$. By dividing both sides of the equations by the baseline risk, R_{00}, we obtain the following null hypothesis equations (1a) written in terms of relative risks, $RR_{ge} = R_{ge}/R_{00}$; this null hypothesis is based on the definition of relative excess risk due to interaction (RERI) (23) that quantifies the magnitude of additive interaction:

$$H_0 : \begin{cases} \mathrm{RERI}_{G=1} = \mathrm{RR}_{11} - \mathrm{RR}_{10} - \mathrm{RR}_{01} + 1 = 0 \\ \mathrm{RERI}_{G=2} = \mathrm{RR}_{21} - \mathrm{RR}_{20} - \mathrm{RR}_{01} + 1 = 0. \end{cases} \tag{1a}$$

It is notable that one major advantage of RERI (versus risk differences themselves) is that it can be estimated using case-control data. That is, under a rare-disease assumption, each relative risk, RR_{ge}, in the above null hypothesis can be approximated by an odds ratio, OR_{ge}:

$$H_0 : \begin{cases} \beta_{G_1 E} = \log\left\{ \dfrac{\exp(\beta_{G_1}) + \exp(\beta_E) - 1}{\exp(\beta_{G_1} + \beta_E)} \right\} (\Longleftrightarrow \mathrm{RERI}_{G=1} = 0) \\ \beta_{G_2 E} = \log\left\{ \dfrac{\exp(\beta_{G_2}) + \exp(\beta_E) - 1}{\exp(\beta_{G_2} + \beta_E)} \right\} (\Longleftrightarrow \mathrm{RERI}_{G=2} = 0), \end{cases} \tag{2a}$$

Table 1. Disease Risk for Each Combination of Gene and Environment Using the Additive Risk Model in Equation 1

	***E* = 0**	***E* = 1**
$G = 0$	$b_0 + \mathbf{x}_i^T\mathbf{b}_\mathbf{x} (= R_{00})$	$b_0 + b_E + \mathbf{x}_i^T\mathbf{b}_\mathbf{x} (= R_{01})$
$G = 1$	$b_0 + b_{G_1} + \mathbf{x}_i^T\mathbf{b}_\mathbf{x} (= R_{10})$	$b_0 + b_{G_1} + b_E + b_{G_1E} + \mathbf{x}_i^T\mathbf{b}_\mathbf{x} (= R_{11})$
$G = 2$	$b_0 + b_{G_2} + \mathbf{x}_i^T\mathbf{b}_\mathbf{x} (= R_{20})$	$b_0 + b_{G_2} + b_E + b_{G_2E} + \mathbf{x}_i^T\mathbf{b}_\mathbf{x} (= R_{21})$

Abbreviations: *E*, environmental factor; *G*, genetic factor; *R*, risk.

which is derived from the following saturated logit model (see Web Table 1, available at https://doi.org/10.1093/aje/kwab124):

$$\mathrm{logit}\{P(D_i = 1|G_i, E_i, \mathbf{x}_i)\} = \beta_0 + \beta_{G_1}G_{1i} + \beta_{G_2}G_{2i} + \beta_E E_i + \beta_{G_1E}G_{1i}E_i + \beta_{G_2E}G_{2i}E_i + \boldsymbol{\beta}_i^T\mathbf{b}_\mathbf{x}. \quad (2)$$

Incorporation of the trend effect of genotype

The derivation and incorporation of the trend effect of genotype (i.e., the linear effect of *G*) in the additive risk model has been introduced previously (9). We note that the "usual" additive coding of genotype for the trend effect in a multiplicative risk model (i.e., coding *G* as 0, 1, or 2 depending on the number of a minor allele and treating *G* as a numeric variable) is not considered here due to the identity link function (versus logit function) used in the additive risk model. Based on the previous results (9), the trend effect of genotype can be expressed as:

$$R_{20} - R_{10} = R_{10} - R_{00} \text{ and } R_{21} - R_{11} = R_{11} - R_{01}.$$

This implies that for each fixed value of *E*, the increment in disease risk for *G* = 1 versus *G* = 0 is equal to the increment for *G* = 2 versus *G* = 1. By dividing both sides of the above equations by R_{00}, we obtain the following trend effect relations expressed in relative risks:

$$\begin{cases} \mathrm{RR}_{20} - 2\mathrm{RR}_{10} + \quad 1 = 0 \\ \mathrm{RR}_{21} - 2\mathrm{RR}_{11} + \mathrm{RR}_{01} = 0, \end{cases} \quad (3a)$$

These equations can be written in terms of the odds ratios in the saturated logit model in equation 2 (Web Table 1):

$$\begin{cases} \beta_{G_2} = \log\left\{2\exp\left(\beta_{G_1}\right) - 1\right\} \\ \beta_{G_2E} = \log\left\{\frac{2\exp\left(\beta_{G_1}+\beta_{G_1E}\right)-1}{\exp\left(\beta_{G_2}\right)}\right\} \end{cases}, \quad (3b)$$

It can be shown that the second equations that involve β_{G_2E} in equations 2a and 3b are identical based on simple algebra. Therefore, the null hypothesis for testing additive interaction under the trend effect of genotype in equation 2a reduces to testing for one RERI parameter:

$$H_0: \beta_{G_1E} = \log\left\{\frac{\exp\left(\beta_{G_1}\right)+\exp\left(\beta_E\right)-1}{\exp\left(\beta_{G_1}+\beta_E\right)}\right\} \;(\Longleftrightarrow \mathrm{RERI}_{G=1} = 0), \quad (3c)$$

To estimate this RERI, we use a maximum likelihood under the trend effect constraints in equation 3b to obtain the estimates for the regression parameters in the logit model in the equation. We use a standard prospective likelihood, $L = \prod_i \Pr(D_i = d_i|G_i, E_i) = \prod_i\{R_i^{d_i}(1 - R_i)^{1-d_i}\}, d_i = 0, 1,$ that does not rely on G-E independence. The RERI can be estimated as follows:

$$\widehat{\mathrm{RERI}}_{\mathrm{UML,tr}} = e^{\hat{\beta}_{G_1,\mathrm{UML,tr}}+\hat{\beta}_{E,\mathrm{UML,tr}}+\hat{\beta}_{G_1E,\mathrm{UML,tr}}} - e^{\hat{\beta}_{G_1,\mathrm{UML,tr}}} - e^{\hat{\beta}_{E,\mathrm{UML,tr}}} + 1,$$

where the parameter estimates are the trend-constrained maximum likelihood estimates from the prospective likelihood (equivalent to the retrospective likelihood without assuming G-E interaction introduced in the next section).

Retrospective likelihood-based inference for G-E independence

To incorporate G-E independence, we employ a retrospective likelihood framework that uses the profile likelihood–based maximum likelihood for estimation (17). The retrospective profile likelihood is given by

$$P(\mathrm{G}, E, \mathbf{X}|D) = \frac{P(D|G, E, \mathbf{X})P(G|E, \mathbf{X})P(E, \mathbf{X})}{\sum_{G,E,\mathbf{x}} P(D|G, E, \mathbf{X})P(G|E, \mathbf{X})P(E, \mathbf{X})},$$

where in the right-hand-side numerator, the first component shows the disease risk given *G*, *E*, and *X*, which can be obtained from equation 2; the second component is given by

$$\mathrm{logit}(P(GE, \mathbf{X})) = \begin{cases} \eta_0 + \eta_1^T X, & \text{with G} - \text{E independence} \\ \eta_0 + \eta_1^T X + \theta E, & \text{without G} - \text{E independence} \end{cases} \quad (4)$$

and the third component $P(E, \mathbf{X})$ is left completely nonparametric. To impose G-E independence, we use the first expression in equation 4—imposing the constraint of $\theta = 0$—which we define as the constrained maximum likelihood method (CML) versus the unconstrained maximum likelihood method (UML) in the second expression in equation 4, which does not use the independence assumption; UML is equivalent to estimating the model parameters under a prospective likelihood (15, 24). To integrate the trend effect of genotype into the retrospective likelihood framework, we estimate the parameters of the logit model (i.e., $\hat{\boldsymbol{\beta}}_{\mathrm{CML,tr}}$) using the profile likelihood implemented in the CGEN package under the constraints shown in equation 3b. Then RERI under G-E independence is estimated as follows:

$$\widehat{\mathrm{RERI}}_{\mathrm{CML,tr}} = e^{\hat{\beta}_{G_1,\mathrm{CML,tr}}+\hat{\beta}_{E,\mathrm{CML,tr}}+\hat{\beta}_{G_1E,\mathrm{CML,tr}}} - e^{\hat{\beta}_{G_1,\mathrm{CML,tr}}} - e^{\hat{\beta}_{E,\mathrm{CML,tr}}} + 1.$$

The variances of $\widehat{\mathrm{RERI}}_{\mathrm{CML,tr}}$ and $\widehat{\mathrm{RERI}}_{\mathrm{UML,tr}}$, $\mathrm{Var}(\widehat{\mathrm{RERI}}_{\mathrm{CML,tr}})$ and $\mathrm{Var}(\mathrm{R\hat{E}RI}_{\mathrm{UML,tr}})$ are estimated using the delta method (25).

Proposed empirical Bayes-type estimator of RERI

We propose a robust empirical Bayes–type estimator of RERI for testing additive interaction under the trend effect of genotype as follows:

$$\begin{aligned}
&\widehat{\mathrm{RERI}}_{\mathrm{EB,tr}}\\
&= \frac{\left(\widehat{\mathrm{RERI}}_{\mathrm{UML,tr}} - \widehat{\mathrm{RERI}}_{\mathrm{CML,tr}}\right)^2}{\left(\widehat{\mathrm{RERI}}_{\mathrm{UML,tr}} - \widehat{\mathrm{RERI}}_{\mathrm{CML,tr}}\right)^2 + \mathrm{Var}\left(\widehat{\mathrm{RERI}}_{\mathrm{UML,tr}}\right)}\\
&\quad \widehat{\mathrm{RERI}}_{\mathrm{UML,tr}}\\
&+ \frac{\mathrm{Var}\left(\widehat{\mathrm{RERI}}_{\mathrm{UML,tr}}\right)}{\left(\widehat{\mathrm{RERI}}_{\mathrm{UML,tr}} - \widehat{\mathrm{RERI}}_{\mathrm{CML,tr}}\right)^2 + \mathrm{Var}\left(\widehat{\mathrm{RERI}}_{\mathrm{UML,tr}}\right)}\\
&\quad \widehat{\mathrm{RERI}}_{\mathrm{CML,tr}} = f\,(a, b, c)\,,
\end{aligned} \tag{5}$$

where $a = \widehat{\mathrm{RERI}}_{\mathrm{UML,tr}}, b = \widehat{\mathrm{RERI}}_{\mathrm{CML,tr}}$ and $c = \mathrm{Var}\left(\widehat{\mathrm{RERI}}_{\mathrm{UML,tr}}\right)$.

The intuition behind the above estimator is that if G-E independence holds, the CML and UML estimators consistently estimate the same true RERI parameter, and hence the difference between the 2 estimates (i.e., $(\widehat{\mathrm{RERI}}_{\mathrm{UML,tr}} - \widehat{\mathrm{RERI}}_{\mathrm{CML,tr}})$) becomes small relative to $\mathrm{Var}(\widehat{\mathrm{RERI}}_{\mathrm{UML,tr}})$. Therefore, $\widehat{\mathrm{RERI}}_{\mathrm{EB,tr}}$ will move towards $\widehat{\mathrm{RERI}}_{\mathrm{CML,tr}}$, and vice versa. Equation 5 can also be expressed as a shrinkage estimator (7, 15) as follows:

$$\widehat{\mathrm{RERI}}_{\mathrm{EB,tr}} = \widehat{\mathrm{RERI}}_{\mathrm{UML,tr}} + K_{\mathrm{tr}}\left(\widehat{\mathrm{RERI}}_{\mathrm{CML,tr}} - \widehat{\mathrm{RERI}}_{\mathrm{UML,tr}}\right), \tag{5a}$$

where K_{tr} is the shrinkage factor that determines the extent by which the UML estimator is shrunk towards the CML estimator and is given by $K_{\mathrm{tr}} = V_{\mathrm{tr}}(V_{\mathrm{tr}} + \delta_{\mathrm{tr}}\delta_{\mathrm{tr}}^T)$, where $\delta_{\mathrm{tr}} = \widehat{\mathrm{RERI}}_{\mathrm{UML,tr}} - \widehat{\mathrm{RERI}}_{\mathrm{CML,tr}}$, and $V_{\mathrm{tr}} = \mathrm{Var}(\widehat{\mathrm{RERI}}_{\mathrm{UML,tr}})$. We derive the variance of the proposed empirical Bayes estimator as:

$$\begin{aligned}
&\mathrm{Var}\left(\widehat{\mathrm{RERI}}_{\mathrm{EB,tr}}\right)\\
&= \mathbf{A}^T \begin{bmatrix} \mathrm{Var}\left(\widehat{\mathrm{RERI}}_{\mathrm{UML,tr}}\right) & \mathrm{Cov}\left(\widehat{\mathrm{RERI}}_{\mathrm{UML,tr}}, \widehat{\mathrm{RERI}}_{\mathrm{CML,tr}}\right) \\ \mathrm{Cov}\left(\widehat{\mathrm{RERI}}_{\mathrm{UML,tr}}, \widehat{\mathrm{RERI}}_{\mathrm{CML,tr}}\right) & \mathrm{Var}\left(\widehat{\mathrm{RERI}}_{\mathrm{CML,tr}}\right) \end{bmatrix} \mathbf{A},
\end{aligned}$$

where

$$\mathbf{A}^T = \begin{bmatrix} \frac{\partial f}{\partial a} & \frac{\partial f}{\partial b} \end{bmatrix}$$

and

$$\begin{aligned}
&\mathrm{Cov}\left(\widehat{\mathrm{RERI}}_{\mathrm{UML,tr}}, \widehat{\mathrm{RERI}}_{\mathrm{CML,tr}}\right) = \left[\frac{\partial \widehat{\mathrm{RERI}}_{\mathrm{UML,tr}}}{\partial \hat{\boldsymbol{\beta}}_{\mathrm{UML,tr}}}\right]^T\\
&\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}_{\mathrm{UML,tr}}, \hat{\boldsymbol{\beta}}_{\mathrm{CML,tr}}\right) \left[\frac{\partial \widehat{\mathrm{RERI}}_{\mathrm{CML,tr}}}{\partial \hat{\boldsymbol{\beta}}_{\mathrm{CML,tr}}}\right].
\end{aligned}$$

The covariance term $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{\mathrm{UML,tr}}, \hat{\boldsymbol{\beta}}_{\mathrm{CML,tr}})$ is computed using the asymptotic covariance matrix

$$\begin{aligned}
&\mathrm{Cov}\begin{pmatrix} \hat{\boldsymbol{\beta}}_{\mathrm{UML,tr}} \\ \hat{\boldsymbol{\beta}}_{\mathrm{CML,tr}} \end{pmatrix} =\\
&\begin{bmatrix} \left(\mathbf{I}^{\mathrm{UML}}\right)^{-1} \mathrm{Var}\left(\sum_i \mathbf{U}_i^{\mathrm{UML}}\right)\left(\mathbf{I}^{\mathrm{UML}}\right)^{-1T} & \left(\mathbf{I}^{\mathrm{UML}}\right)^{-1} \mathrm{Cov}\left(\sum_i \mathbf{U}_i^{\mathrm{UML}}, \sum_i \mathbf{U}_i^{\mathrm{CML}}\right) \left(\mathbf{I}^{\mathrm{CML}}\right)^{-1T} \\ \left(\mathbf{I}^{\mathrm{CML}}\right)^{-1} \mathrm{Cov}\left(\sum_i \mathbf{U}_i^{\mathrm{CML}}, \sum_i \mathbf{U}_i^{\mathrm{UML}}\right) \left(\mathbf{I}^{\mathrm{UML}}\right)^{-1T} & \left(\mathbf{I}^{\mathrm{CML}}\right)^{-1} \mathrm{Var}\left(\sum_i \mathbf{U}_i^{\mathrm{CML}}\right)\left(\mathbf{I}^{\mathrm{CML}}\right)^{-1T} \end{bmatrix},
\end{aligned}$$

where $\mathbf{I}$ denotes the information matrix, and $\mathbf{U}_i$ denotes individual score functions. A detailed derivation of $\mathrm{Var}(\widehat{\mathrm{RERI}}_{\mathrm{EB,tr}})$ is shown in the Web Appendix 1. The resulting Wald test for the proposed method is as follows:

$$Z_{\mathrm{EB,tr}} = \widehat{\mathrm{RERI}}_{\mathrm{EB,tr}} / \sqrt{\mathrm{Var}\left(\widehat{\mathrm{RERI}}_{\mathrm{EB,tr}}\right)} \sim N\left(0, 1\right).$$

Simulation study

We conducted simulation studies to assess the performance of the proposed empirical Bayes–type estimator under the trend effect of genotype (i.e., empirical Bayes–trend test), comparing with existing methods. In the first set of simulations, we evaluated the type I error, bias, and mean squared error of the proposed test under varying magnitudes of the departure from the G-E independence assumption, comparing with 2 existing methods—the additive interaction test under the trend effect of genotype using the retrospective likelihood (i.e., CML-trend test) that relies on G-E independence and the standard additive interaction test under the trend effect of genotype without the independence assumption (i.e., UML-trend test). In the second set of simulations, we assessed the power of the proposed empirical Bayes–trend test versus the existing methods under varying magnitudes of additive interaction measured as RERI, assuming G-E independence. For each simulation, the minor allele frequency (MAF) of the genetic factor G was varied as $p_g = 0.3, 0.1, 0.05$, with the following genotype probability for $G = 0$, 1, and 2 respectively: $(1 - P_g)^2, 2P_g(1 - P_g)$, and $P_g{}^2$. The prevalence of the environmental factor E was assumed to be $p_e = 0.2$. We fixed the marginal odds ratio (MOR) for G (MOR(G))—that is, the disease relative risk for $G = 1$ (versus $G = 0$) if E is ignored in the analysis—at 1.1, reflecting the modest strength of association typically observed in genome-wide association studies (GWAS). We varied the MOR for E (MOR(E)) from 1.5 to 2.5, 3, and 3.5. The expressions for MOR are given in Web Appendix 2. For each type I error simulation, we generated 10,000 cases and 10,000 controls, whereas for power simulation we considered 5,000 cases and 5,000 controls. We additionally considered an alternative case-control ratio of 3:7 (i.e., 6,000 cases and 14,000 controls) for type I error simulation to examine whether a different ratio might have an impact on the relative performance of the methods. We assumed a disease prevalence of 0.01 to reflect the rare-disease assumption. The saturated logit model shown in equation 2 (without the covariate terms, $\boldsymbol{\beta}_i^T \mathbf{b_x}$) was used to simulate data. The parameter values for β_0, β_{G_1}, β_{G_2}, β_{E}, β_{G_1E}, and β_{G_2E} in the logit model were chosen to meet the given values of MOR(G), MOR(E), and RERI for each scenario (see Web Table 2 for type 1 error, bias, and mean squared error simulation, and Web Table 3 for power simulation). For simulating the departure from G-E independence (i.e., varying degrees G-E dependence), we used the following relation for modeling the correlation between gene and environment among controls:

$$\operatorname{logit}(P(E = 1)) = \theta_0 + \theta_{GE}G,$$

with $\theta_0 = \log(0.427)$ and the value of θ_{GE} ranging from $\pm\log(1.01)$ to $\pm\log(2)$ (see Web Table 4), considering both positive and negative departures from G-E independence that were chosen to reflect the empirical findings for exposure-SNP associations (26). For each set of simulations under a given value for θ_{GE}, 50,000 replicates were generated for type I error evaluation. Type I error was estimated as the proportion of the replicates that led to a P value higher than a given significance level α. For power simulation without the violation of G-E independence, RERI values of 0.8, 1, 1.2 were used to vary the magnitude of additive interaction using 1,000 replicates.

Late-onset Alzheimer disease data

We applied the proposed empirical Bayes–trend test to examine interactions between SNPs and the ε4 allele of the *APOE* gene (i.e., *APOE*-ε4) using GWAS data for LOAD. In particular, we were interested in evaluating the performance of the proposed method when G-E independence (or G-G independence for SNP × *APOE*-ε4 interaction analysis) is known to be violated. The *APOE* gene is located at chromosome 19, and *APOE*-ε4 has been shown to be strongly associated with LOAD risk (27). The SNPs located in the *APOE* gene and their neighboring SNPs in chromosome 19 are in high linkage disequilibrium (i.e., correlated) with *APOE*-ε4, hence violating the G-G independence assumption. Given that the goal of this analysis is to assess the performance of the proposed method under the violation of G-G independence, we focused on a total of 59,661 SNPs belonging to chromosome 19 that are present in the GWAS data. This data contained the genotypes of ∼5 million SNPs for 8,861 cases and 7,613 controls who are of northwestern European ancestry, collected from 18 different studies, and were made available by the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site and similar LOAD repositories (see Web Appendix 3 and Web Tables 5–7). The *APOE*-ε4 variable was coded as 1 for mutation carriers versus 0 for noncarriers, as commonly used (28, 29). The model was adjusted for age, sex, 3 principal components for population stratification, and a study variable. For the CML- and empirical Bayes–trend tests, the study variable was used as a stratification variable. In addition to the chromosome 19 analysis, we also conducted the analysis of GWAS data across all chromosomes to compare the performance of the tests in terms of type I error (see Web Appendix 3).

Lung cancer data

We also applied the proposed method to investigate gene × smoking interactions using GWAS data for lung cancer. The lung cancer data set contains data for 5,739 cases and 5,848 controls from 4 studies—the Environment and Genetics in Lung Cancer Etiology study (30), the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (31), the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (32), and the Cancer Prevention Study II Nutrition Cohort (33). This data includes 15 SNPs that were identified from previous GWAS ($P < 5 \times 10^{-8}$), conducted in either European (34–37) or Asian populations (32–34) listed in the National Human Genome Research Institute GWAS catalog (https://www.ebi.ac.uk/gwas/) (see Web Table 8 for information on these SNPs) (38). The smoking variable was coded as 1 for ever-smokers and 0 for never-smokers. We adjusted for age, sex, and study variable, where the study was used as a stratification variable for the retrospective empirical Bayes analysis.

RESULTS

Simulation results

Figure 1 shows the results of type I error simulation under varying degrees of departure from G-E independence (the *x*-axis) and significance levels (different panels). As expected, the CML-trend test (the dotted curves in Figure 1), assuming G-E independence, shows substantially increased type I error rates as the extent of the violation of the assumption increases (i.e., the correlation between gene and environment

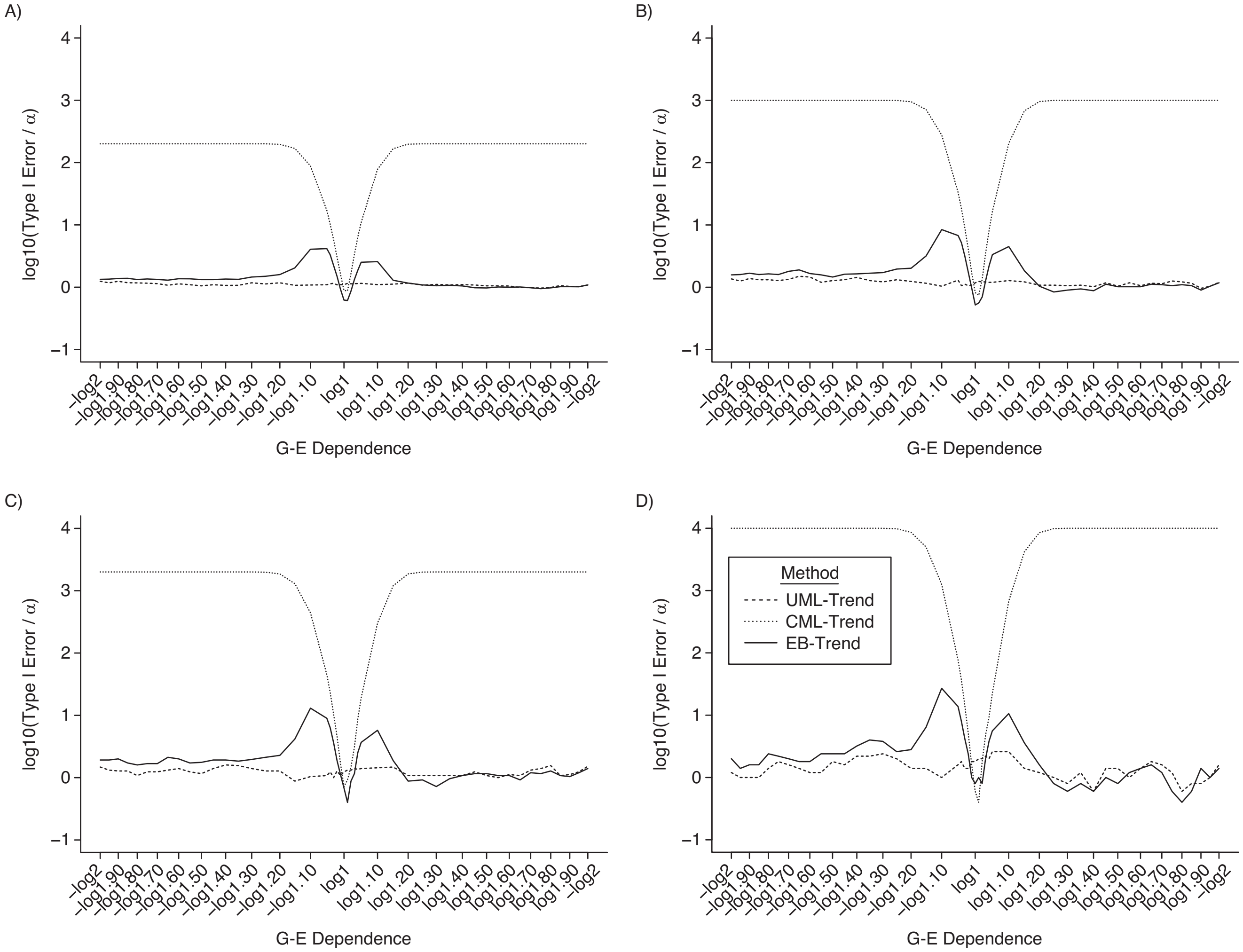

Figure 1. Type I error simulation under varying departure from gene-environment (G-E) independence (θ_{GE},the *x*-axis) across different type I error thresholds: A) α = 0.005; B) α = 0.001; C) α = 0.0005; D) α = 0.0001. The *y*-axis shows the log (to the base 10) of the ratio between an observed type I error rate and α (i.e., log10(observed type I error/α)). Three additive interaction tests—unconstrained maximum likelihood (UML)-trend, constrained maximum likelihood (CML)-trend, and empirical Bayes (EB)-trend tests—are applied to simulated data sets generated under the null hypothesis (i.e., relative excess risk due to interaction = 0); 50,000 replicated data sets are simulated for 10,000 cases and 10,000 controls with minor allele frequency (MAF) = 0.3, marginal odds ratio (MOR)(G) = 1.1, MOR(E) = 1.5. The parameters used for this simulation are presented in Web Table 2. The results of the simulation under different MAF (0.1,0.05), MOR(E) (2.5,3,3.5), and case-control ratio are shown in Web Figure 1 (different MAF values), Web Figure 2 (different MOR(E) values), and Web Figure 3 (different case-control ratio).

increases in the underlying population). On the other hand, the proposed empirical Bayes–trend test (the dashed curves in Figure 1)—using a robust shrinkage estimator that takes into account the uncertainty around G-E independence—shows reduced type I error rates, which are close to the levels obtained under the standard UML-trend test without assuming G-E independence. The inflation of type I error using the CML-trend test becomes more severe under a more stringent significance threshold (Figure 1D, with α = 0.0001) compared with those under larger thresholds (α = 0.0005, 0.001, and 0.005). Similar results are observed under the alternative MAF values, MOR(E) values, and case-control ratio (Web Figures 1–3).

Figure 2 shows the bias and mean squared error for the 3 methods (UML-trend, CML-trend and empirical Bayes–trend tests) under varying levels of departure from G-E independence. As the magnitude of the departure increases, the bias using the CML-trend test increases dramatically compared with the empirical Bayes–trend test and the UML-trend test. Similarly, the CML-trend test shows increased mean squared error when the assumption is violated, whereas the proposed empirical Bayes–trend method demonstrates lower mean squared error that is close to the level of the UML-trend test. Consistent results are observed under the alternative MAF and MOR(E) values (Web Figures 4–6).

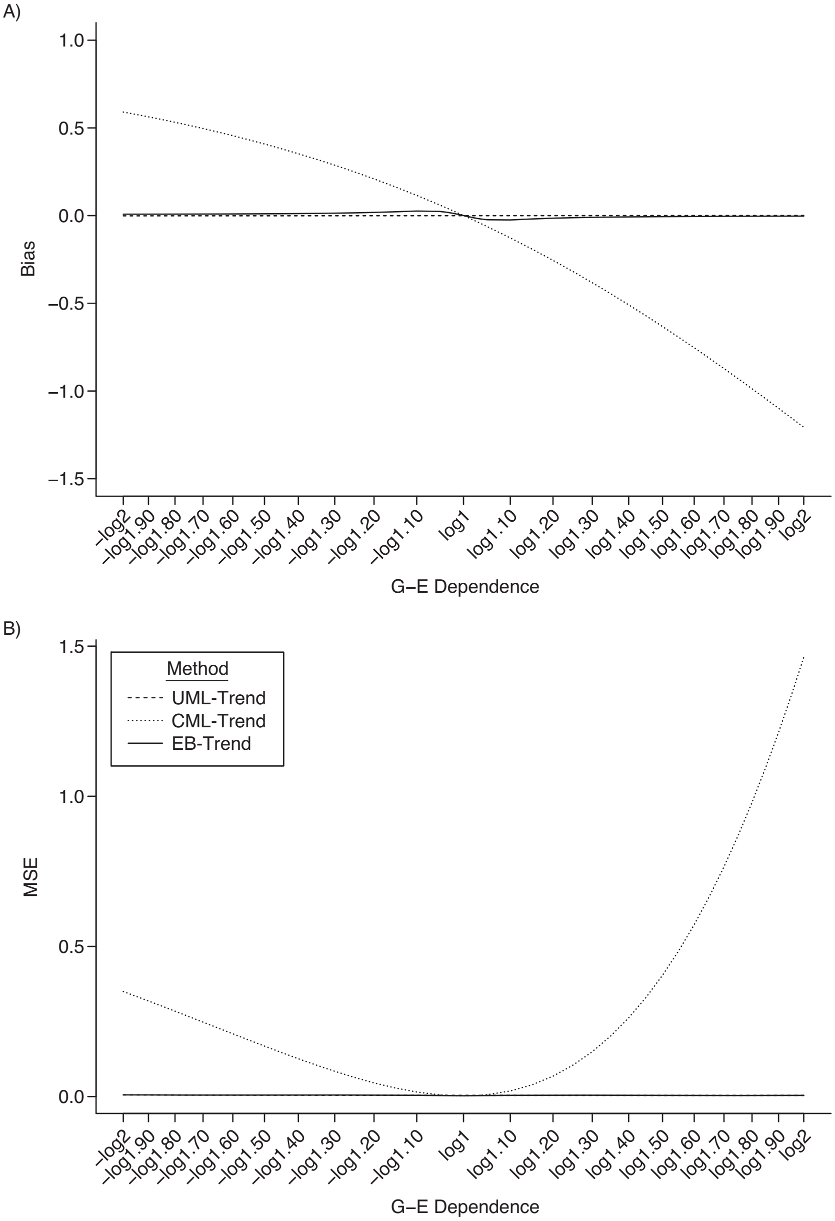

Figure 2. Simulation for evaluating bias (A) and mean squared error (MSE) (B) of relative excess risk due to interaction (RERI) under varying magnitude of gene-environment (G-E) dependence (θ_{GE},the *x*-axis) that includes the negative correlation and positive correlation. The simulated data sets are generated under the null hypothesis (RERI = 0), and 50,000 replicated data sets were simulated for 10,000 cases and 10,000 controls with marginal odds ratio (MOR) = 0.3, marginal odds ratio (MOR)(G) = 1.1, MOR(E) = 1.5. The parameters used for this simulation are presented in Web Table 2**.** The results of the simulation under different MAF (0.1,0.05) and MOR(E) (2.5,3,3.5) are shown in Web Figure 4 (different MAFs for the given MOR(E) = 1.5), Web Figure 5 (different MOR(E) and MAF values for evaluating bias), and Web Figure 6 (different MOR(E) and MAF values for evaluating MSE). EB, empirical Bayes; CML, constrained maximum likelihood; UML, unconstrained maximum likelihood.

Figure 3 shows the results of power simulations under the assumption of G-E independence. As expected, the CML-trend shows the largest power when this assumption is met, and the proposed empirical Bayes–trend test shows a higher power compared with the standard UML-trend but lower than the CML-trend test. We observed similar results under the alternative MAFs and MOR(E) values (Web Figure 7).

Analysis for SNP × *APOE*-ε4 interaction for LOAD

The results of SNP × *APOE*-ε4 interaction analysis for LOAD are shown in Figure 4, applying the existing additive CML-trend test (Figure 4A) and the proposed additive empirical Bayes–trend test (Figure 4B) on chromosome 19, which harbors the *APOE* gene. As we expected, the result

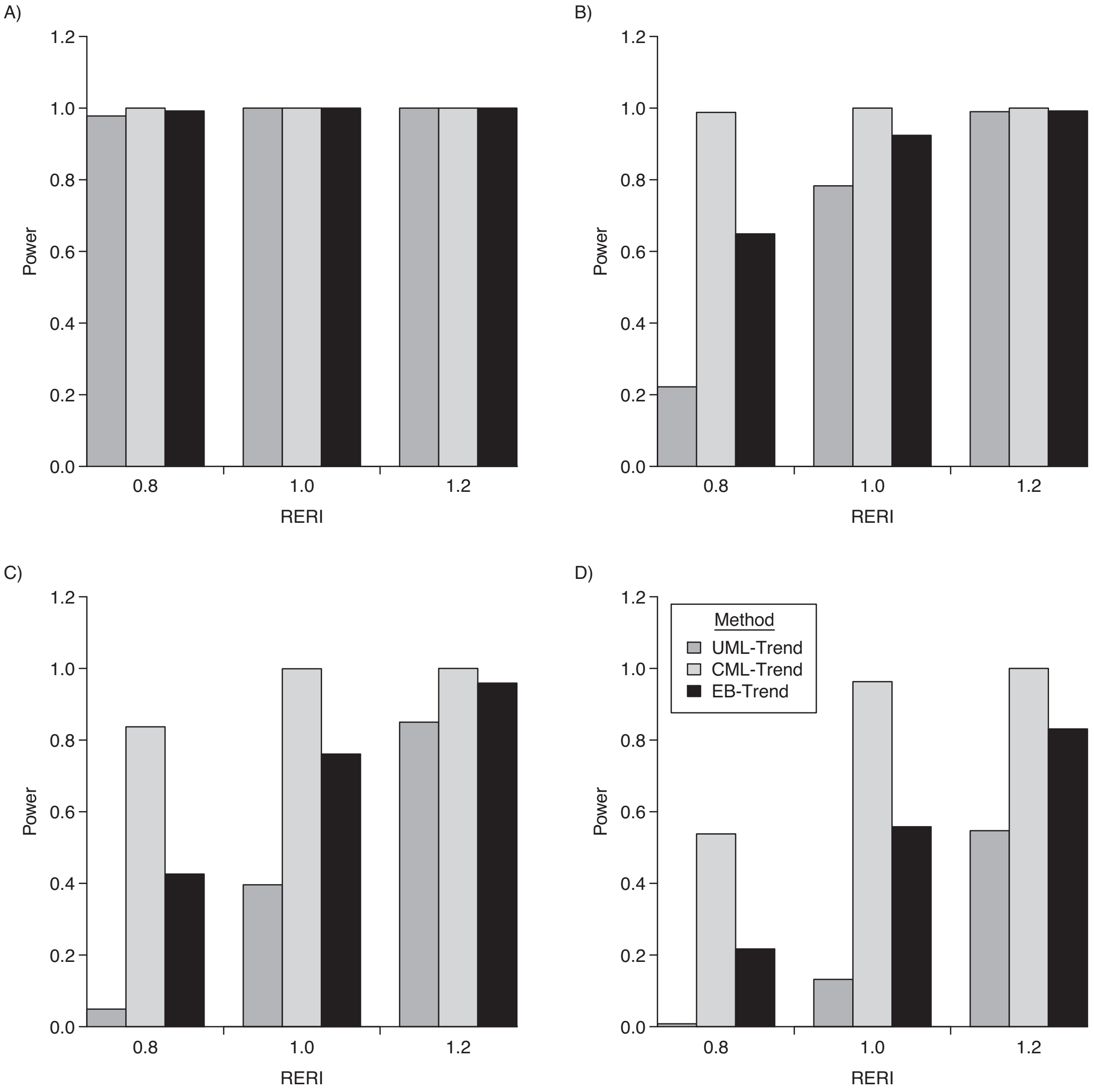

Figure 3. Power comparison of the 3 tests—unconstrained maximum likelihood (UML)-trend, constrained maximum likelihood (CML)-trend, and empirical Bayes (EB)-trend tests (based on the α threshold of 1×10^{-8})—under varying magnitudes of additive interaction (relative excess risk due to interaction, RERI) and varying marginal odds ratio (MOR)(E). A) MOR(E) = 1.5; B) MOR(E) = 2.5; C) MOR(E) = 3; D) MOR(E) = 3.5; 1,000 replicated data sets are simulated for 5,000 cases and 5,000 controls with marginal odds ratio (MOR) = 0.3, MOR(G) = 1.1. The parameters used for this simulation are presented in Web Table 3. The results of the simulation under different MAF (0.1,0.05) values are shown in Web Figure 7.

based on the CML-trend test, assuming G-G independence, shows a high peak of false-positive signals near the base-pair positions between 45,048,120 and 45,868,038, where the *APOE* is located; 349 of a total of 747 SNPs in this region exceed the genome-wide significance level ($P < 5 \times 10^{-8}$) using the CML-trend test. On the other hand, none of the SNPs exceed the level using the proposed empirical Bayes test, where the false positive signals disappear, demonstrating the robustness of the proposed method. The results under the UML-trend are shown in Web Figure 8. The analysis results using GWAS data across all chromosomes are presented in Web Figures 9 and 10. Overall, the results show that the CML-trend test has substantial inflation in type I error with a noticeable departure from the 45-degree line in the quantile-quantile plot, whereas the empirical Bayes–trend and UML-trend tests show reasonably well-controlled type I error across the entire genome.

Analysis for SNP × smoking interaction for lung cancer data analysis

The results for SNP × smoking interaction analysis, shown in Table 2, demonstrate that 3 SNPs on 15q25.1, rs8034191 (empirical Bayes–trend test, $P = 2.317 \times 10^{-12}$), rs1051730

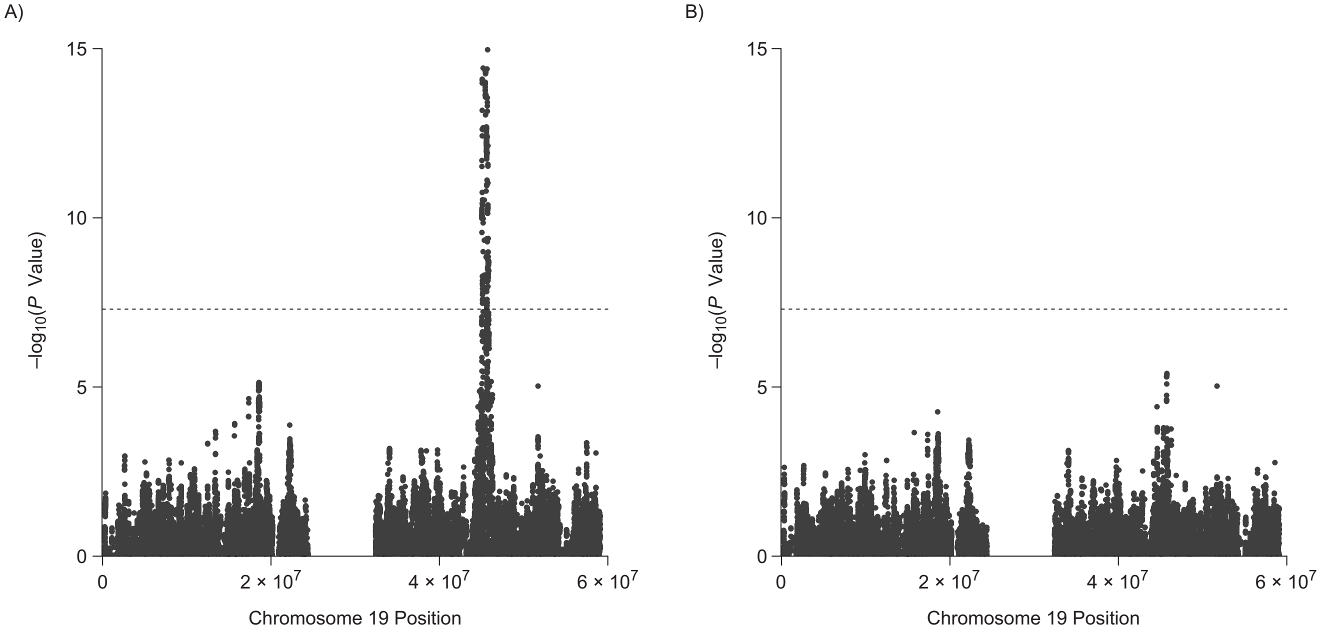

Figure 4. Manhattan plots for single nucleotide polymorphism (SNP) × apolipoprotein E (*APOE*)-ε4 interaction analysis for chromosome 19, which harbors the apolipoprotein E gene, using data from 8,861 cases and 7,613 controls of northwestern European ancestry, collected from 18 different studies. The *y*-axis shows $-\log_{10}(P)$ values from testing SNP × APOE-ε4 interaction using: A) the additive constrained maximum likelihood (CML)-trend test; B) the additive empirical Bayes (EB)-trend test (the proposed test). The result using the additive unconstrained maximum likelihood (UML)-trend test is shown in Web Figure 8. The *y*-axis is truncated at 15. The dashed line corresponds to genome-wide significance level of $P = -\log_{10}(5 \times 10^{-8})$.

(empirical Bayes–trend test, $P = 7.158 \times 10^{-13}$), and rs8042374 (empirical Bayes–trend test, $P = 6.393 \times 10^{-9}$) are statistically significant ($P < 0.05/45 = 0.001$ based on the Bonferroni correction). We note that highly significant additive interactions of rs8034191 ($P = 2 \times 10^{-10}$) and rs1051730 ($P = 1 \times 10^{-9}$) with smoking intensity (number of cigarettes smoked per day) were previously reported in the literature (39). The RERI estimates, 95% confidence intervals, and stratified odds ratios by smoking status for the top 3 SNPs are shown in Web Table 9. The results of testing for the G-E independence assumption using the controls data (i.e., SNP-smoking association among controls) are shown in Web Table 10, where none of the 15 SNPs show a significant departure from the assumption. Other notable SNPs are rs31489 (empirical Bayes–trend test, $P = 0.006$) on 5p15.33 and rs3117582 (empirical Bayes–trend test, $P = 0.007$) on 6p21.33, which approached but did not reach the statistical significance threshold.

DISCUSSION

We have proposed a robust test for evaluating additive interaction under the trend effect of genotype using an empirical Bayes-type shrinkage estimator. Simulation study under varying levels of departures from G-E independence shows that the proposed method is robust against the violation of G-E independence while providing an adequate balance between bias and efficiency versus the existing methods.

It is notable that the top 2 SNPs on 15q25.1 (rs1051730 and rs8034191) that showed significant interactions with smoking intensity on lung cancer risk (39) are also reported to be associated with nicotine dependence (26, 40–45), thus potentially violating the SNP-smoking independence assumption. However, our tests for assessing G-E independence did not show any significant departures. This could be due to the potential lack of power of the G-E association test or to the difference in environmental exposures used across different analyses (smoking initiation/status versus smoking intensity).

To the best of our knowledge, the proposed method is the first approach that evaluates additive G × E interaction under the trend effect of genotype by combining the traditional prospective likelihood-based and the retrospective likelihood-based estimators to relax the strict G-E independence assumption. The application of the proposed empirical Bayes–trend method to examine SNP × *APOE*-ε4 interaction for LOAD demonstrated a substantial improvement in controlling false-positive results over the existing CML-trend method when the assumption on the G-E (or G-G) independence was violated. Finally, we implemented the proposed method in the CGEN package (46), which enables its wide application for identifying various interactions among researchers in genetic epidemiology.

Our study is not without limitations. The proposed method is based on a Wald test, which tends to have a lower power compared with a likelihood ratio test, especially for alternatives sufficiently far from the null value or for additive

Table 2. Single Nucleotide Polymorphism × Smoking Interaction Analysis for Lung Cancer (5,739 Cases and 5,848 Controls)[a] Among Populations With European or Asian Ancestry

SNP	Region	Mapped Genes	*P* Value[b]			RERI (SE)[b]		
			UML-Trend	CML-Trend	EB-Trend	UML-Trend	CML-Trend	EB-Trend
rs1051730	15q25.1	*CHRNA3*	5.346×10^{-12}[c]	7.222×10^{-13}[c]	7.158×10^{-13}[c]	2.472 (0.359)	2.500 (0.348)	2.500 (0.348)
rs8034191	15q25.1	*AGPHD1*	9.678×10^{-12}[c]	2.318×10^{-12}[c]	2.317×10^{-12}[c]	2.164 (0.318)	2.160 (0.308)	2.160 (0.308)
rs8042374	15q25.1	*CHRNA3*	1.79×10^{-18}[c]	1.815×10^{-9}[c]	6.393×10^{-9}[c]	−1.870 (0.332)	−1.708 (0.284)	−1.739 (0.300)
rs31489	5p15.33	*CLPTM1L*	0.03	0.001[c]	0.006	−0.562 (0.259)	−0.746 (0.234)	−0.684 (0.248)
rs3117582	6p21.33	*BAG6*; *APOM*	0.005	0.006	0.007	1.361 (0.483)	1.130 (0.414)	1.173 (0.434)
rs2395185	6p21.32		0.048	0.021	0.021	−0.678 (0.343)	−0.651 (0.282)	−0.651 (0.282)
rs4975616	5p15.33	*MIR4457—CLPTM1L*	0.088	0.007	0.022	−0.425 (0.249)	−0.599 (0.223)	−0.542 (0.237)
rs401681	5p15.33	*CLPTM1L*	0.092	0.005	0.028	−0.416 (0.247)	−0.646 (0.229)	−0.539 (0.245)
rs2736100	5p15.33	*TERT*	0.274	0.063	0.1	−0.224 (0.204)	−0.341 (0.183)	−0.312 (0.190)
rs3817963	6p21.32	*BTNL2*	0.378	0.125	0.156	−0.314 (0.355)	−0.459 (0.299)	−0.438 (0.308)
rs4324798	6p22.1	*NOP56P1—RPL13P*	0.067	0.273	0.177	0.966 (0.527)	0.475 (0.433)	0.703 (0.521)
rs7216064	17q24.2	*BPTF*	0.31	0.547	0.493	0.328 (0.323)	0.172 (0.286)	0.202 (0.294)
rs9387478	6q22.1	*ROS1—DCBLD1*	0.439	0.968	0.768	−0.190 (0.245)	0.008 (0.204)	−0.070 (0.237)
rs753955	13q12.12	*TNFRSF19—MTCO3P2*	0.964	0.769	0.777	−0.011 (0.251)	−0.065 (0.222)	−0.063 (0.222)
rs10937405	3q28	*TP63*	0.89	0.984	0.982	0.031 (0.225)	0.004 (0.195)	0.004 (0.195)

Abbreviations: AGPHD1, aminoglycoside phosphotransferase domain-containing protein 1; APOM, apolipoprotein M; BAG6, BAG cochaperone 6; BPTF, bromodomain PHD finger transcription factor; BTNL2, butyrophilin like 2; CHRNA3, cholinergic receptor nicotinic alpha 3 subunit; CLPTM1L, cleft lip and palate transmembrane 1 like; CML, constrained maximum likelihood; DCBLD1, discoidin, CUB and LCCL domain containing 1; EB, empirical Bayes; MIR4457, microRNA 4457; MTCO3P2, MT-CO3 pseudogene 2; NOP56P1, NOP56 ribonucleoprotein pseudogene 1; RERI, relative excess risk due to interaction; ROS1, ROS proto-oncogene 1, receptor tyrosine kinase; RPL13P, ribosomal protein L13 pseudogene; SE, standard error; SNP, single nucleotide polymorphism; TERT, telomerase reverse transcriptase; TNFRSF19, TNF receptor superfamily member 19; TP63, tumor protein p63; UML, unconstrained maximum likelihood.

[a] The lung cancer data set contains data for 5,739 cases and 5,848 controls from 4 studies (30–33).

[b] The results are presented for the EB-trend, CML-trend, and UML-trend tests of SNP × smoking interaction. The rows are sorted by the *P* values of EB-trend test. RERI denotes $RERI_{G=1}$ for trend model.

[c] Statistically significant *P* values, $\alpha = 0.0011$ (=0.05/(15 × 3)) was applied for statistical significance.

risk models (47–49). In addition, the proposed method is not directly applicable to analyzing imputed genotype data. This is because the constraints for additive joint effects of gene and environment and the trend effect of genotype are ascertained in terms of a categorical genotype variable, which cannot be defined for continuous imputed dosage scores. We also note that the proposed empirical Bayes–trend test can have modest bias in the presence of G-E dependence. The UML-trend test is only guaranteed to provide unbiased estimates in the presence of G-E correlation when there is no exposure misclassification (50, 51). For evaluating gene-smoking interaction for lung cancer analysis, we focused on smoking initiation (ever vs. never). It is possible that the analysis based on smoking intensity could be different from the results presented in this study, especially in terms of the relative performance of the empirical Bayes– versus CML-trend tests. In our simulation study, we used a range of MAF values between 5% and 30%, but we did not consider a rare MAF below 5%. This was partly due to the heavy computational burden involved in conducting large simulations under rare MAF values; the proposed algorithm requires a minimum of 5 samples for each combination of gene and environment categories, and hence a much larger sample size is needed to run simulations under rare MAF values (e.g., 1%) than the sample size we considered (10,000 cases and 10,000 controls). A further evaluation with potentially smaller values of MAF would be also helpful, and that will require more intense computations.

In conclusion, we developed an empirical Bayes-type shrinkage estimator for testing G × E interaction under an additive risk model that incorporates the trend effect of genotype. Future research directions include an extension of the proposed method to incorporate imputed genotype data and a potential modification of the data-adaptive approach using a likelihood ratio test to improve power.

ACKNOWLEDGMENTS

Author affiliations: Quantitative Sciences Unit, Department of Medicine, Stanford University School of Medicine, Stanford University, Stanford, California, United States (Nilotpal Sanyal, Matthieu de Rochemonteix, Summer S. Han); Department of Neurology, Stanford University School of Medicine, Stanford University, Stanford, California, United States (Valerio Napolioni, Michaël E. Belloy, Michael D. Greicius); Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, United States (Neil E. Caporaso, Maria Teresa Landi); Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States (Nilanjan Chatterjee); Stanford Cancer Institute, Stanford University School of Medicine, Stanford University, Stanford, California, United States (Summer S. Han); and Department of Neurosurgery, Stanford University School of Medicine, Stanford University, Stanford, California, United States (Summer S. Han).

This work was funded by the National Cancer Institute (grant 1R37CA226081 to S.S.H.) and National Institutes of Health (grants AG066515 and AG060747 to M.D.G.). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (National Institutes of Health grant U01 AG024904 and Department of Defense award number W81XWH-12-2-0012). Alzheimer's Disease Neuroimaging Initiative is funded by the National Institute on Aging and the National Institute of Biomedical Imaging and Bioengineering.

Data availability: All data used in the Alzheimer disease genetics analyses are available upon application to: dbGaP (https://www.ncbi.nlm.nih.gov/gap/), NIAGADS (https://www.niagads.org/), LONI (https://ida.loni.usc.edu/), Synapse (https://www.synapse.org/), Rush (https://www.radc.rush.edu/), or NACC (https://naccdata.org/). All data used for SNP x smoking interaction analysis for lung cancer are available upon application to: dbGaP (https://www.ncbi.nlm.nih.gov/gap/).

Biological samples used in this study were stored at principal investigators' institutions and at the National Cell Repository for Alzheimer's Disease (NCRAD) at Indiana University, which receives government support under a cooperative agreement (grant U24 AG21886) awarded by the National Institute on Aging (NIA). 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; Alzheimer's Disease Genetics Consortium (ADGC), U01 AG032984, RC2 AG036528; NACC, U01 AG016976; NIA-LOAD (Columbia University), U24 AG026395, U24 AG026390, R01AG041797; Banner Sun Health Research Institute P30 AG019610; Boston University, P30 AG013846, U01 AG10483, R01 CA129769, R01 MH080295, R01 AG017173, R01 AG025259, R01 AG048927, R01AG33193, R01 AG009029; Columbia University, P50 AG008702, R37 AG015473, R01 AG037212, R01 AG028786; Duke University, P30 AG028377, AG05128; Group Health Research Institute, UO1 AG006781, UO1 HG004610, UO1 HG006375, U01 HG008657; Indiana University, P30 AG10133, R01 AG009956, RC2 AG036650; Johns Hopkins University, P50 AG005146, R01 AG020688; Massachusetts General Hospital, P50 AG005134; Mayo Clinic, P50 AG016574, R01 AG032990, KL2 RR024151; Mount Sinai School of Medicine, P50 AG005138, P01 AG002219; New York University, P30 AG08051, UL1 RR029893, 5R01AG012101, 5R01AG022374, 5R01AG013616, 1RC2AG036502, 1R01AG035137; Northwestern University, P30 AG013854; Oregon Health & Science University, P30 AG008017, R01 AG026916; Rush University, P30 AG010161, R01 AG019085, R01 AG15819, R01 AG17917, R01 AG030146, R01 AG01101, RC2 AG036650, R01 AG22018; TGEN, R01 NS059873; University of Alabama at Birmingham, P50 AG016582, UL1RR02777; University of Arizona, R01 AG031581; University of California, Davis, P30 AG010129; University of California, Irvine, P50 AG016573, P50 AG016575, P50 AG016576, P50 AG016577; University of California, Los

Angeles, P50 AG016570; University of California, San Diego, P50 AG005131; University of California, San Francisco, P50 AG023501, P01 AG019724; University of Kentucky, P30 AG028383, AG05144; University of Michigan, P30 AG053760 and AG063760; University of Pennsylvania, P30 AG010124; University of Pittsburgh, P50 AG005133, AG030653, AG041718, AG07562, AG02365; University of Southern California, P50 AG005142; University of Texas Southwestern, P30 AG012300; University of Miami, R01 AG027944, AG010491, AG027944, AG021547, AG019757; University of Washington, P50 AG005136, R01 AG042437; University of Wisconsin, P50 AG033514; Vanderbilt University, R01 AG019085; and Washington University, P50 AG005681, P01 AG03991, P01 AG026276.

ROSMAP study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by National Institute on Aging grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, U01AG46152, the Illinois Department of Public Health, and the Translational Genomics Research Institute.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu/). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Conflict of interest: none declared.

REFERENCES

1. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. *Stat Med*. 1983;2(2):243–251.
2. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. *Am J Epidemiol*. 1980;112(4):467–470.
3. Walter SD, Holford TR. Additive, multiplicative, and other models for disease risks. *Am J Epidemiol*. 1978;108(5): 341–346.
4. Han SS, Chatterjee N. Review of statistical methods for gene-environment interaction analysis. *Current Epidemiology Reports*. 2018;5(1):39–45.
5. Han SS, Rosenberg PS, Garcia-Closas M, et al. Likelihood ratio test for detecting gene (G)-environment (E) interactions under an additive risk model exploiting G-E independence for case-control data. *Am J Epidemiol*. 2012;176(11): 1060–1067.
6. Kim S, Wang M, Tyrer JP, et al. A comprehensive gene–environment interaction analysis in ovarian cancer using genome-wide significant common variants. *Int J Cancer*. 2019;144(9):2192–2205.
7. Liu G, Mukherjee B, Lee S, et al. Robust tests for additive gene-environment interaction in case-control studies using gene-environment independence. *Am J Epidemiol*. 2018; 187(2):366–377.
8. Ni A, Satagopan JM. Estimating additive interaction effect in stratified two-phase case-control design. *Hum Hered*. 2019; 84(2):90–108.
9. de Rochemonteix M, Napolioni V, Sanyal N, et al. A likelihood ratio test for gene-environment interaction based on the trend effect of genotype under an additive risk model using the gene-environment independence assumption. *Am J Epidemiol*. 2021;190(1):129–141.
10. Tchetgen Tchetgen EJ, Shi X, Wong BHW, et al. A general approach to detect gene (G)-environment (E) additive interaction leveraging G-E independence in case-control studies. *Stat Med*. 2019;38(24):4841–4853.
11. Garcia-Closas M, Rothman N, Figueroa JD, et al. Common genetic polymorphisms modify the effect of smoking on absolute risk of bladder cancer. *Cancer Res*. 2013;73(7): 2211–2220.
12. Thompson WD. Effect modification and the limits of biological inference from epidemiologic data. *J Clin Epidemiol*. 1991;44(3):221–232.
13. Siemiatycki J, Thomas DC. Biological models and statistical interactions: an example from multistage carcinogenesis. *Int J Epidemiol*. 1981;10(4):383–387.
14. Umbach DM, Weinberg CR. Designing and analysing case-control studies to exploit independence of genotype and exposure. *Stat Med*. 1997;16(15):1731–1743.
15. Chen Y-H, Chatterjee N, Carroll RJ. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *J Am Stat Assoc*. 2009;104(485):220–233.
16. Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of case–control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*. 2008;64(3): 685–694.
17. Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*. 2005; 92(2):399–418.
18. Han SS, Rosenberg PS, Ghosh A, et al. An exposure-weighted score test for genetic associations integrating environmental risk factors. *Biometrics*. 2015; 71(3):596–605.
19. Chu H, Nie L, Cole SR. Estimating the relative excess risk due to interaction: a Bayesian approach. *Epidemiology*. 2011; 22(2):242–248.
20. Nie L, Chu H, Li F, et al. Relative excess risk due to interaction: resampling-based confidence intervals. *Epidemiology*. 2010;21(4):552–556.
21. Richardson DB, Kaufman JS. Estimation of the relative excess risk due to interaction and associated confidence bounds. *Am J Epidemiol*. 2009;169(6):756–760.
22. Albert PS, Ratnasinghe D, Tangrea J, et al. Limitations of the case-only design for identifying gene-environment interactions. *Am J Epidemiol*. 2001;154(8):687–693.
23. Hosmer DW, Lemeshow S. Confidence interval estimation of interaction. *Epidemiology*. 1992;3(5):452–456.
24. Zhao LP, Li SS, Khalid N. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *Am J Hum Genet*. 2003;72(5):1231–1250.
25. Wasserman L. *All of Statistics: A Concise Course in Statistical Inference*. New York, NY: Springer; 2013.
26. Saccone NL, Emery LS, Sofer T, et al. Genome-wide association study of heavy smoking and daily/nondaily smoking in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *Nicotine Tob Res*. 2018;20(4): 448–457.

27. Corder EH, Saunders AM, Strittmatter WJ, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*. 1993;261(5123): 921–923.
28. Jun G, Ibrahim-Verbaas CA, Vronskaya M, et al. A novel Alzheimer disease locus located near the gene encoding tau protein. *Mol Psychiatry*. 2016;21(1):108–117.
29. Jun GR, Chung J, Mez J, et al. Transethnic genome-wide scan identifies novel Alzheimer's disease loci. *Alzheimers Dement*. 2017;13(7):727–738.
30. Landi MT, Chatterjee N, Yu K, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet*. 2009;85(5):679–691.
31. Group ACPS. The Alpha-Tocopherol, Beta-Carotene Lung Cancer Prevention Study: design, methods, participant characteristics, and compliance. *Ann Epidemiol*. 1994;4(1): 1–10.
32. Purdue MP, Mink PJ, Hartge P, et al. Hormone replacement therapy, reproductive history, and colorectal adenomas: data from the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial (United States). *Cancer Causes Control*. 2005;16(8):965–973.
33. Calle EE, Rodriguez C, Jacobs EJ, et al. The American Cancer Society Cancer Prevention Study II nutrition cohort: rationale, study design, and baseline characteristics. *Cancer*. 2002;94(9):2490–2501.
34. Amos CI, Wu X, Broderick P, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet*. 2008;40(5):616–622.
35. Hung RJ, McKay JD, Gaborieau V, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*. 2008;452(7187): 633–637.
36. McKay JD, Hung RJ, Gaborieau V, et al. Lung cancer susceptibility locus at 5p15. *Nat Genet*. 2008;40(12): 1404–1406.
37. Wang Y, Broderick P, Webb E, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet*. 2008; 40(12):1407–1409.
38. Saccone NL, Culverhouse RC, Schwantes-An T-H, et al. Multiple independent loci at chromosome 15q25. 1 affect smoking quantity: a meta-analysis and comparison with lung cancer and COPD. *PLoS Genet*. 2010;6(8):e1001053.
39. VanderWeele TJ, Asomaning K, Tchetgen Tchetgen EJ, et al. Genetic variants on 15q25.1, smoking, and lung cancer: an assessment of mediation and interaction. *Am J Epidemiol*. 2012;175(10):1013–1020.
40. Brazel DM, Jiang Y, Hughey JM, et al. Exome chip meta-analysis fine maps causal variants and elucidates the genetic architecture of rare coding variants in smoking and alcohol use. *Biol Psychiatry*. 2019;85(11):946–955.
41. Erzurumluoglu AM, Liu M, Jackson VE, et al. Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci. *Mol Psychiatry*. 2020; 25(10):2392–2409.
42. Liu JZ, Tozzi F, Waterworth DM, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet*. 2010;42(5):436–440.
43. Thorgeirsson TE, Geller F, Sulem P, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*. 2008;452(7187):638–642.
44. Thorgeirsson TE, Gudbjartsson DF, Surakka I, et al. Sequence variants at *CHRNB3–CHRNA6* and *CYP2A6* affect smoking behavior. *Nat Genet*. 2010;42(5):448–453.
45. Furberg H, Kim Y, Dackor J, et al. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet*. 2010;42(5):441.
46. Bhattacharjee S, Chatterjee N, Han S, et al. CGEN: An R package for analysis of case-control studies in genetic epidemiology, version. 3.28.0. https://bioconductor.org/packages/release/bioc/html/CGEN.html. Accessed July 9, 2021.
47. Fears TR, Benichou J, Gail MH. A reminder of the fallibility of the Wald statistic. *Am Stat*. 1996;50(3):226–227.
48. Hauck WW Jr, Donner A. Wald's test as applied to hypotheses in logit analysis. *J Am Stat Assoc*. 1977;72(360a): 851–853.
49. Storer BE, Wacholder S, Breslow NE. Maximum likelihood fitting of general risk models to stratified data. *J R Stat Soc Ser C Appl Stat*. 1983;32(2):172–181.
50. Dudbridge F, Fletcher O. Gene-environment dependence creates spurious gene-environment interaction. *Am J Hum Genet*. 2014;95(3):301–307.
51. Lindström S, Yen Y-C, Spiegelman D, et al. The impact of gene-environment dependence and misclassification in genetic association studies incorporating gene-environment interactions. *Hum Hered*. 2009;68(3):171–181.