American Journal of Epidemiology
© Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health 2020.
This work is written by (a) US Government employee(s) and is in the public domain in the US.
Vol. 190, No. 1
DOI: 10.1093/aje/kwaa132
Advance Access publication:
September 1, 2020

Practice of Epidemiology

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

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

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

Initially submitted December 19, 2019; accepted for publication June 30, 2020.

Several statistical methods have been proposed for testing gene-environment (G-E) interactions under additive risk models using data from genome-wide association studies. However, these approaches have strong assumptions from underlying genetic models, such as dominant or recessive effects that are known to be less robust when the true genetic model is unknown. We aimed to develop a robust trend test employing a likelihood ratio test for detecting G-E interaction under an additive risk model, while incorporating the G-E independence assumption to increase power. We used a constrained likelihood to impose 2 sets of constraints for: 1) the linear trend effect of genotype and 2) the additive joint effects of gene and environment. To incorporate the G-E independence assumption, a retrospective likelihood was used versus a standard prospective likelihood. Numerical investigation suggests that the proposed tests are more powerful than tests assuming dominant, recessive, or general models under various parameter settings and under both likelihoods. Incorporation of the independence assumption enhances efficiency by 2.5-fold. We applied the proposed methods to examine the gene-smoking interaction for lung cancer and gene–apolipoprotein E ε4 interaction for Alzheimer disease, which identified 2 interactions between apolipoprotein E ε4 and loci membrane-spanning 4-domains subfamily A (*MS4A*) and bridging integrator 1 (*BIN1*) genes at genome-wide significance that were replicated using independent data.

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

Abbreviations: APOE, apolipoprotein E; BIN1, bridging integrator 1; G-E, gene-environment; GWAS, genome-wide association study; LOAD, late-onset Alzheimer disease; LRT, likelihood ratio test; LRT-P, prospective likelihood ratio test; LRT-R, retrospective likelihood ratio test; MOR, marginal odds ratio; RERI, relative excess risk due to interaction; SNP, single nucleotide polymorphism.

A gene-environment interaction is defined as the joint effect of genetic and environmental factors that cannot be explained by their separate marginal effects (1). Statistically, an interaction is defined as a departure from the underlying disease risk model, which depends on the selection of proper scale for measuring the presence of interactions (2). While multiplicative interaction based on a logit model is more commonly used for case-control studies, additive interaction has been shown to be more relevant for evaluating prevention or intervention strategies in public health decision-making (3–5). Additive interaction measures the departure from an additive risk model, which assumes that gene and environment act additively on the risk of the disease itself (6, 7). It has been shown that conceptual models for biological interactions translate to the presence of interaction on the additive scale and not necessarily the multiplicative scale (8). Recently, several studies reported G × G or G × E interaction findings applying additive interaction tests for various complex diseases (9–11).

A number of methods have been proposed for detecting additive interactions between risk factors for case-control studies (12–16), including the methods for estimating the

Downloaded from https://academic.oup.com/aje/article/190/1/129/5900107 by University of Southern California user on 10 June 2021

confidence intervals of the relative excess risk due to interaction (RERI), which measures the magnitude of additive interaction (17–19). A likelihood ratio test (LRT) has been developed for additive interaction based on a set of constraints to ensure the additive joint effects of gene and environment, which also incorporates the gene-environment (G-E) independence assumption to increase power (13). Another recent study extended this method and applied an empirical Bayes-type approach to exploit the G-E independence assumption in a data-adaptive way (14). However, the main limitation of these methods is their strong assumptions from the underlying genetic model, such as dominant or recessive effects that are known to be less robust when the true genetic model is unknown. A more general genetic model that treats genotype as a categorical variable (i.e., allowing 2 separate parameters for the main effect of genotype) has also been proposed (13), but it has been shown to have reduced power due to increased degrees of freedom (df). It is also known that most common variants associated with complex diseases from genome-wide association study (GWAS) have been identified under an additive genetic model (20), where the effect of genotype is linear (i.e., the trend effect of genotype), rather than the effect being under a dominant or recessive model. Studies have shown that association tests based on the trend effect of genotype are more robust compared with dominant or recessive model–based tests under varying true genetic models (21), which have been widely used in analyzing the main effects of common variants in most recent GWAS (22, 23).

In this study, we developed a new statistical test for detecting additive interaction that incorporates the trend effect of genotype by extending the previously developed LRT (13). We use a constrained likelihood approach to impose 2 sets of constraints for: 1) the trend effect of genotype and 2) the additive joint effect of gene and environment, using the parameter estimates obtained from a saturated logit model. To incorporate the G-E independence assumption, a retrospective likelihood is used. We conducted a simulation study with varying genetic models and varying magnitudes of interaction. We applied the proposed method to examine the gene-smoking interaction for lung cancer and both gene–apolipoprotein E (*APOE*) ε4 and gene-sex interactions for late-onset Alzheimer disease (LOAD). We implemented this method in the freely available R (R Foundation for Statistical Computing, Vienna, Austria) package, CGEN (https://www.bioconductor.org/packages/release/bioc/html/CGEN.html).

METHODS

Additive interaction under a dominant or recessive genetic model

For subject i, G_i is a binary genetic factor, where $G_i = 1$ if the subject has at least 1 copy of the variant of interest in a single nucleotide polymorphism (SNP) and $G_i = 0$ otherwise for a dominant genetic model; $G_i = 1$ if the subject has 2 copies of the variant in a SNP and $G_i = 0$ otherwise for a recessive genetic model. E_i is an environmental risk factor that is a categorical variable, and D_i is the disease status ($D_i = 1$ if one has a disease and $D_i = 0$ otherwise). Without loss of generality, we assume a binary E_i in this setting. An additive risk model assumes that G_i and E_i act additively on the risk of the disease itself without any (nonidentity) link function: $P(D_i = 1|G_i, E_i) = b_0 + b_G G_i + b_E E_i$. A departure from this model is called an additive interaction, which can be tested using the null hypothesis of $H_0 : b_{GE} = 0$ in the following model:

$$R_i = P(D_i = 1|G_i, E_i) = b_0 + b_G G_i + b_E E_i + b_{GE} G_i E_i.$$

Suppose $R_{ge} = P(D = 1|G = g, E = e)$ for $g, e = 0, 1$. The null hypothesis, $H_0 : b_{GE} = 0$, can be expressed as $R_{11} - R_{01} = R_{10} - R_{00}$, which implies that the risk difference associated with G_i is constant across different levels of E_i. By dividing this equation by the reference risk R_{00}($RR_{ij} = R_{ij}/R_{00}$), we obtain the relative risk relationship for the null: $RR_{11} - RR_{01} = RR_{10} - 1$. Assuming a rare disease, we can approximate a relative risk by an odds ratio, and hence we obtain $H_0 : OR_{11} - OR_{01} = OR_{10} - 1$. This implies that additive interaction can be tested using the following null hypothesis:

$$H_0 : \exp(\beta_G + \beta_E + \beta_{GE}) - \exp(\beta_G) - \exp(\beta_E) + 1 = 0, \quad (1)$$

where the parameters are estimated using the saturated logit model,

$$\text{logit}\{\Pr(D_i = 1|G_i, E_i)\} = \beta_0 + \beta_G G_i + \beta_E E_i + \beta_{GE} G_i E_i. \quad (2)$$

Additive interaction can be tested using a likelihood ratio test using the null hypothesis in equation 1 against the alternative using the model in equation 2, which gives a 1-df test. The magnitude of additive interaction can be measured by RERI, which is defined as: $RERI_G = \exp(\beta_G + \beta_E + \beta_{GE}) - \exp(\beta_G) - \exp(\beta_E) + 1$. This additive interaction LRT can be generalized to G_i as a 3-level categorical variable (G_i= 0, 1, 2) and multilevel E_i ($E_i = e_1, e_2, \ldots, e_k$), which has been derived in a previous study (13) and implemented in CGEN (24). This general model–based test has larger df (df = 2k versus df = k for a dominant or recessive model–based test).

Additive interaction under the trend effect of genotype (additive genetic model)

To extend the method described in the previous section to an approach based on the trend effect (linear effect) of genotype, we consider the following saturated additive model for a 3-category genetic factor, G_i (G_i= 0, 1, 2):

$$R_i = P(D_i = 1|G_i, E_i) = b_0 + b_{G_1} G_{1i} + b_{G_2} G_{2i} + b_E E_i + b_{G_1 E} G_{1i} E_i + b_{G_2 E} G_{2i} E_i,$$

Where G_{1i} and G_{2i} are dummy variables indicating whether subject i has 1 or 2 copies of the variant in a given SNP. Any covariates can also be included in the model without

Downloaded from https://academic.oup.com/aje/article/190/1/129/5900107 by University of Southern California user on 10 June 2021

Table 1. Disease Risk and Odds Ratio for Gene (G = 0,1,2) and Environment (E = 0,1) Based on the Saturated Additive Risk Model and the Saturated Logistic Regression Model, Respectively

G	Disease Risk		Odds Ratio	
	$E=0$	$E=1$	$E=0$	$E=1$
$G=0$	$R_{00}(=b_0)$	$R_{01}(=b_0+b_E)$	$OR_{00}(=1)$	$OR_{01}(=\exp(\beta_E))$
$G=1$	$R_{10}(=b_0+b_{G_1})$	$R_{11}(=b_0+b_{G_1}+b_E+b_{G_1E})$	$OR_{10}(=\exp(\beta_{G_1}))$	$OR_{11}(=\exp(\beta_{G_1}+\beta_E+\beta_{G_1,E}))$
$G=2$	$R_{20}(=b_0+b_{G_2})$	$R_{21}(=b_0+b_{G_2}+b_E+b_{G_2E})$	$OR_{20}(=\exp(\beta_{G_2}))$	$OR_{21}(=\exp(\beta_{G_2}+\beta_E+\beta_{G_2,E}))$

Abbreviations: E, environmental factor; G, genetic factor; OR, odds ratio; R, risk.

affecting the derivations that are shown in this section. The corresponding risk and odds ratio for each G and E are shown in Table 1. The trend effect of G on the disease risk (or linear effect of G) can be expressed as:

$$R_{20}-R_{10}=R_{10}-R_{00}\ \text{(for } E=0\text{) and} \tag{3}$$

$$R_{21}-R_{11}=R_{11}-R_{01}\ \text{(for } E=1\text{).} \tag{4}$$

The null hypothesis of no additive interaction, $H_0: b_{G_1,E}=b_{G_2,E}=0$, can equivalently be expressed as:

$$H_0: R_{11}-R_{01}=R_{10}-R_{00} \tag{5}$$

and

$$R_{21}-R_{11}=R_{20}-R_{10}. \tag{6}$$

These equations can be rewritten in terms of relative risks by dividing them by the baseline risk, R_{00}. These relative risks can be approximated using odds ratios under a rare disease assumption. We use the following saturated logit model to obtain the odds ratio for each combination of levels of G and E:

$$\begin{aligned}&\text{logit}\{\Pr(D_i=1|G_i,E_i)\}\\&=\beta_0+\beta_{G_1}G_{1i}+\beta_{G_2}G_{2i}+\beta_E E_i\\&\quad+\beta_{G_1E}G_{1i}E_i+\beta_{G_2E}G_{2i}E_i,\end{aligned} \tag{7}$$

which are shown in the last 2 columns in Table 1. Using these odds ratios, the trend effect relationships in equation 3 and equation 4 are expressed as:

$$\begin{aligned}&\beta_{G_2}=\log(2\exp(\beta_{G_1})-1)\\&\quad(\Longleftrightarrow \exp(\beta_{G_2})-\exp(\beta_{G_1})=\exp(\beta_{G_1})-1)\end{aligned} \tag{3b}$$

and

$$\begin{aligned}\beta_{G_2,E}&=\log\left(\frac{2\exp\left(\beta_{G_1}+\beta_{G_1,E}\right)-1}{2\exp\left(\beta_{G_1}\right)-1}\right)\\&=\log\left(\frac{2\exp\left(\beta_{G_1}+\beta_{G_1,E}\right)-1}{\exp\left(\beta_{G_2}\right)}\right).\end{aligned} \tag{4b}$$

We note that the "usual" additive coding of the genotype for the trend effect for multiplicative interaction (i.e., coding G as 0, 1, or 2 depending on the number of a minor allele) is different from the trend effect of genotype for additive interaction (shown in equations 3b and 4b) due to the different link functions used for these models (logit versus identity). Similarly, the null hypothesis in equation 5 and equation 6 is expressed as:

$$\begin{aligned}H_0:\beta_{G_1,E}&=\log\left(\frac{\exp\left(\beta_{G_1}\right)+\exp\left(\beta_E\right)-1}{\exp\left(\beta_{G_1}+\beta_E\right)}\right)\\&(\Longleftrightarrow\ RERI_{G=1}=0)\end{aligned} \tag{5b}$$

and

$$\begin{aligned}\beta_{G_2,E}&=\log\left(\frac{\exp\left(\beta_{G_2}\right)+\exp\left(\beta_E\right)-1}{\exp\left(\beta_{G_2}+\beta_E\right)}\right)\\&(\Longleftrightarrow\ RERI_{G=2}=0).\end{aligned} \tag{6b}$$

Based on simple algebra, we show that the equations in equation 5b and equation 6b are identical when the trend effect relation equations in equation 3b and equation 4b hold. Therefore, under the trend effect of genotype, the null hypothesis for testing additive interaction is equation 5b.

Constrained LRT for additive interaction under the trend effect of genotype

In conducting an LRT, we first fit the saturated logit model shown in equation 7 with the 2 trend effect–related constraints shown in equation 3b and equation 4b (i.e., the full model). This model includes the following free parameters to be estimated by the maximum likelihood method: $\beta_0,\beta_{G_1},\beta_E$, and $\beta_{G_1,E}$. In the second step, we further impose the additional null hypothesis related constraint from equation 5b to fit the null model, which has 3 free parameters: β_0,β_{G_1}, and β_E. We conduct an LRT by comparing the full model with the null model, which gives a 1-df test. For fitting these models, we use a "prospective likelihood" that is commonly used for case-control data: $L=\prod_i \Pr(D_i=d|G_i,E_i)=\prod_i\{R_i^{d_i}(1-R_i)^{1-d_i}\}$ (for $d_i=0,1$), and henceforth we refer to the corresponding test as LRT-P.

Downloaded from https://academic.oup.com/aje/article/190/1/129/5900107 by University of Southern California user on 10 June 2021

Extension to incorporate the G-E independence assumption

For case-control analysis, an assumption of G-E independence in the underlying population can be used to enhance the power of G-E tests (13, 25, 26). It has been shown that a retrospective likelihood for case-control data can help exploit the G-E independence assumption to obtain efficient parameter estimates of a general logistic model (25). More recently, a study showed how the retrospective likelihood framework can enable the G–E independence assumption to enhance the power of additive interaction tests (13).

To incorporate the G-E independence assumption in our method, we used the profile likelihood-based approach developed by Chatterjee and Carroll (25) to develop a retrospective LRT for additive interaction under the trend effect of genotype, henceforth denoted LRT-R. The profile likelihood is given by:

$$L = \prod_i \Pr(D_i = d, G_i = g | E_i = e, S_i = s, R = 1),$$

where R indicates whether or not a subject is included in the case-control design, and S_i is a stratifying variable (such as ethnicity or principle components of population stratification).

Simulation study

We conducted a simulation study to evaluate the type I error rates and power of the proposed additive interaction tests under the trend effect of genotype and compared the power with those under the existing dominant, recessive, and general model–based additive interaction tests. For the power simulation, we considered the 4 different scenarios where the true disease risk model is under: 1) the trend effect of genotype (i.e., additive genetic model), 2) the dominant model, 3) the recessive model, and 4) the general model (i.e., a departure from the dominant, recessive, and trend effect models). In each scenario, we varied the magnitude of additive interaction (i.e. $RERI_{G=1)}$ from 0.8 to 1 and 1.2. We assumed that G and E are independently distributed in the underlying population. The minor allele frequency of G was set as 0.3, and the prevalence for E was set as $\Pr(E_i = 1) = 0.2$. We also assumed the disease was rare, so that disease-free subjects approximately represented the underlying population. The disease prevalence was assumed to be 0.01. We fixed the marginal odds ratio (MOR) for the gene (MOR(G))—that is, the disease odds ratio for $G = 1$ (vs. $G = 0$) if environment is ignored in the analysis—at 1.25, reflecting the modest strength of genetic association typically observed in GWAS, and varied the MOR for environment (MOR(E)) from 2.5 to 3.5. The saturated logit model in equation 7 was used to simulate data. We chose the parameter values for $\beta_0, \beta_{G_1}, \beta_{G_2}, \beta_E$, $\beta_{G_1,E}$, and $\beta_{G_2,E}$ in the logistic model so that MOR(G), MOR(E), and RERI are fixed at given values (see Table 2). To compare the power of the proposed trend effect–based LRT-R versus LRT-P, we used the parameters under the truth of the trend effect model (the first 6 rows in Table 2). In each simulation, we generated G and E data for 5,000 cases and 5,000 controls. A significance level of $\alpha = 1.00 \times 10^{-7}$ was used for power calculation with 1,000 replicates, and 10,000 replicates were used for assessing type I error.

Data description for lung cancer and smoking

We applied the proposed method to examine the gene × smoking interaction for lung cancer, where smoking is a known risk factor. We used National Cancer Institute GWAS data that included 5,739 cases and 5,848 controls from 4 studies (27). This data included 14 SNPs that were identified from previous GWAS ($P < 5 \times 10^{-8}$), conducted in either White (28–31) or Asian populations (32–34) listed in the National Human Genome Research Institute GWAS catalog (https://www.ebi.ac.uk/gwas/). The genetic regions for this data include 15q25.1, which is known to interact with smoking under an additive model (16) and is associated with smoking behaviors (35–41). Therefore, we applied only LRT-P to the SNPs in this region. The goal of this analysis was to compare the results using the proposed tests with those using the existing additive interaction tests based on dominant or general models. For each SNP, we applied both LRT-P and LRT-R (except 15q25.1), and the model adjusted for age, gender, and study, where the study variable was used as a stratification variable (S) for LRT-R.

Data description for LOAD and gene × sex and gene × gene interactions

We also applied the proposed methods to the GWAS data for LOAD to examine gene × sex interaction and gene × *APOE* ε4 gene interaction. The ε4 allele of the apolipoprotein E gene (*APOE*, on chromosome 19q13.32, Online Mendelian Inheritance in Man (OMIM)∗107741) is the strongest genetic risk factor for LOAD (42). Individuals carrying 2 copies of the *APOE* ε4 allele have more than a 10-fold increased risk of LOAD (43). It is reported that women have a higher risk of LOAD than men (44). Our goal was to examine interactions between SNP × *APOE* ε4 and SNP × sex by applying the proposed methods. For the discovery phase, we used GWAS data that includes 8,861 cases and 7,613 controls who are of Northwestern European ancestry, collected from 18 different studies. These data were made available by the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site and similar LOAD repositories (Web Appendix 1; Web Tables 1–2) (45). The GWAS data included 89,936 SNPs across the genome, which were pruned from approximately 5 million SNPs using the following procedure: 1) we ranked SNPs by the significance of the main effect for SNP-LOAD risk association; 2) starting from the top ranked SNP, we evaluated pairwise linkage disequilibrium between the given SNP and each other SNP to remove the SNPs that were in high linkage disequilibrium ($r^2 > 0.9$); and 3) we removed SNPs with a minor allele frequency below 5% (45). This procedure was used to reduce the computational burden of the analysis through the removal of highly correlated SNPs but retain the SNPs that were relevant to LOAD risk (45). The pruning was conducted based on PriorityPruner, v.0.1.4 (http://prioritypruner.sourceforge.net/). For the validation phase, we used an independent data set obtained from the second

Downloaded from https://academic.oup.com/aje/article/190/1/129/5900107 by University of Southern California user on 10 June 2021

Table 2. Twenty-Four Sets of Parameter Values Used to Simulate Gene, Environment, and Disease Indicator Using the Logistic Regression Model in Equation 7

MOR and Parameter	Trend Model			Dominant Model			Recessive Model			General Model		
	RERI[a]			RERI[a]			RERI[b]			RERI[a]		
	0.8	1	1.2	0.8	1	1.2	0.8	1	1.2	0.8	1	1.2
MOR(E) = 2.5, MOR(G) = 1.25												
$\exp(\beta_0)$	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
$\exp(\beta_{G1})$	1.160	1.108	1.064	1.160	1.114	1.070	1.000	1.000	1.000	1.160	1.111	1.067
$\exp(\beta_{G2})$	1.320	1.216	1.128	1.160	1.114	1.070	1.172	1.132	1.093	1.237	1.164	1.099
$\exp(\beta_E)$	2.200	2.040	1.880	2.260	2.120	2.030	2.490	2.477	2.440	2.230	2.080	1.954
$\exp(\beta_{G1E})$	1.238	1.393	1.572	1.228	1.369	1.519	1.000	1.000	1.000	1.233	1.381	1.545
$\exp(\beta_{G2E})$	1.419	1.716	2.079	1.228	1.369	1.519	1.186	1.287	1.400	1.320	1.533	1.777
MOR(E) = 3, MOR(G) = 1.25												
$\exp(\beta_0)$	0.007	0.007	0.007	0.006	0.006	0.006	0.007	0.007	0.007	0.007	0.007	0.007
$\exp(\beta_{G1})$	1.186	1.140	1.099	1.189	1.143	1.095	1.000	1.000	1.000	1.187	1.141	1.097
$\exp(\beta_{G2})$	1.372	1.280	1.198	1.189	1.143	1.095	1.199	1.157	1.117	1.277	1.210	1.145
$\exp(\beta_E)$	2.790	2.630	2.450	2.840	2.700	2.520	3.000	2.980	2.960	2.815	2.665	2.485
$\exp(\beta_{G1E})$	1.141	1.257	1.392	1.134	1.245	1.383	1.000	1.000	1.000	1.138	1.251	1.387
$\exp(\beta_{G2E})$	1.244	1.459	1.720	1.134	1.245	1.383	1.112	1.200	1.294	1.188	1.348	1.542

Abbreviations: E, environmental factor; G, genetic factor; MOR, marginal odds ratio; RERI, relative excess risk due to interaction.

[a] RERI denotes $RERI_{G=1}$ for trend, dominant, and general model; the values for $RERI_{G=2}$ for the general model are: 1.175, 1.45, and 1.76, with corresponding $RERI_{G=1}$ values of 0.8, 1, and 1.2, respectively.

[b] RERI denotes $RERI_{G=2}$ for recessive model.

wave of National Institute on Aging—Alzheimer's Disease Centers GWAS (ADC4–7) (Web Appendix 1), which included 1,907 cases and 1,677 controls who were of European ancestry, and comprised the top 2 SNPs that showed significance ($P < 5 \times 10^{-8}$) in our discovery phase (show below). The *APOE* ε4 variable was coded as 1 for mutation carriers versus 0 for noncarriers, as commonly used in the literature (46, 47). This gene is located on chromosome 19q, and hence the SNPs in this gene (and neighboring SNPs) might violate the assumption for G-E (in this case G-G) independence. Therefore, we excluded this region for LRT-R-based SNP × *APOE* ε4 analysis. The model adjusted for age, sex, and study, and we used the study variable as a stratification variable for LRT-R.

RESULTS

Simulation study

The results in Table 3 show that the proposed methods for additive interaction have correct type I error rates for both LRT-P and LRT-R. The power simulation results under the prospective likelihood are shown in Figure 1, where the first column shows the power of the 4 additive interaction tests applied to the data generated under an additive genetic model. The results demonstrate that the proposed trend effect–based interaction test has a larger power compared with the existing (dominant, recessive, and general model–based) tests across different magnitudes of additive interaction (RERI). When data were simulated under the truth of the dominant model (the second column in Figure 1), the additive interaction test assuming the dominant model was most powerful, as expected, and the general model–based and the proposed trend effect–based tests showed comparable power. For data simulated under the model that departs from the dominant and the trend effect models (the third column in Figure 1), the proposed test showed a larger power than the dominant model–based test and had a comparable power to the general model–based test.

We also compared the power of the proposed trend effect–based additive test under the prospective versus retrospective likelihood that assumes the G-E independence assumption. The results in Figure 2 show that the trend effect–based interaction test under the retrospective likelihood provides a larger power than the one under the prospective likelihood across different RERIs. The third column in Figure 2 shows that the retrospective likelihood-based approach has 2.4–2.5 times increased efficiency compared with the prospective likelihood approach, by exploiting the G-E independence assumption. For comparison, the simulation results using

Downloaded from https://academic.oup.com/aje/article/190/1/129/5900107 by University of Southern California user on 10 June 2021

Table 3. Type 1 Error Rates of the Additive Interaction Tests Based on the Trend Effect of the Genotype, Dominant Model, Recessive Model, and General Model

α	LRT-R				LRT-P			
	Trend Test	Dominant Test	Recessive Test	General Test	Trend Test	Dominant Test	Recessive Test	General Test
0.1000	0.0987	0.1007	0.1018	0.1018	0.0987	0.0991	0.0987	0.0997
0.0500	0.0466	0.0497	0.0534	0.0507	0.0497	0.0510	0.0508	0.0494
0.0100	0.0109	0.0090	0.0112	0.0093	0.0076	0.0092	0.0087	0.0097
0.0050	0.0050	0.0047	0.0054	0.0044	0.0034	0.0046	0.0047	0.0051
0.0010	0.0009	0.0009	0.0010	0.0010	0.0009	0.0012	0.0012	0.0011

Abbreviations: LRT-P, prospective likelihood ratio test; LRT-R, retrospective likelihood ratio test.

the multiplicative interaction tests (trend, dominant, recessive, and general) under both likelihoods are shown in Web Figures 1–2. These showed consistent results compared with the additive interaction tests (Figure 1 and Figure 2), where the trend effect–based test showed robustness across different underlying genetic models, and the approach under the retrospective likelihood increased efficiency compared with the prospective likelihood.

Application to the lung cancer and smoking data

The results of the SNP × smoking interaction analysis for lung cancer are shown in Table 4. Among the 14 SNPs we examined, 3 SNPs (rs8034191 and rs8042374 in 15q25.1 and rs31489 in 5p15.33) showed statistical significance (α = 0.0006 based on Bonferroni correction). Web Table 3 displays the 3 × 2 odds ratio tables and RERI estimates for these SNPs. The topmost significant interaction was with rs8034191 ($P = 5.55 \times 10^{-16}$ using the trend effect–based test; RERI = 2.16, 95% CI: 1.54, 2.78), which confirmed the previously reported additive interaction between this SNP and smoking (16). For this SNP, the proposed trend effect–based test showed improved significance ($P = 5.55 \times 10^{-16}$) compared with the dominant ($P = 1.13 \times 10^{-14}$) or general model–based tests ($P = 7.88 \times 10^{-15}$). Increase of risk induced by the C allele of this SNP is significantly larger among ever smokers versus never smokers. Another SNP that reached statistical significance was rs31489 at the cleft lip and palate transmembrane protein 1-like protein gene (*CLPTM1L*) at 5p15.33, which showed improved significance for trend effect–based tests compared with the dominant or general model–based tests, with highest significance observed for the trend effect–based test under the retrospective likelihood ($P = 4.11 \times 10^{-4}$; RERI = −0.75, 95% CI: −1.20, −0.28).

Application to the LOAD GWAS data

Web Table 4 shows the top 30 most significant findings from the SNP × *APOE* ε4 interaction analysis (discovery phase) and is sorted by the significance of the additive interaction test based on the trend effect of genotype under the retrospective likelihood approach. Applying the genome-wide significance level ($\alpha = 5 \times 10^{-8}$), 2 SNPs were statistically significant across the different tests conducted (see Table 5 for odds ratios and RERI estimates), the proposed trend effect–based additive interaction test showing the most significant *P* values for both of them. The topmost significant SNP × *APOE* ε4 interaction was found with rs6733839 on 2q14.3 (additive interaction trend test $P = 5.56 \times 10^{-14}$; RERI = 1.11, 95% CI: 0.788, 1.422), where the absolute increase of risk induced by the T allele was significantly larger among *APOE* ε4 carriers versus noncarriers. To examine this region further, we pulled out all the SNP data in the region harboring rs6733839 (+/− 200kb) and conducted the same set of tests. The zoom-in Manhattan plot in Web Figure 3A shows a peak around the SNP rs6733839, where a neighboring SNP rs4663105 that shows stronger significance ($P = 4.77 \times 10^{-15}$) is highly correlated with rs6733839 ($r^2 = 0.94$). The second-most significant interaction was identified with rs1582763 ($P = 5.32 \times 10^{-9}$; RERI = −0.64, 95% CI: −0.852, −0.426) in 11q12.2, an intergenic variant near membrane-spanning 4-domains subfamily A member 4A (*MS4A4A*). The zoom-in plot for this SNP, shown in Web Figure 3B, shows a peak around rs1582763. The results of SNP × sex interaction did not show any significant findings ($P < 5 \times 10^{-8}$) (see Web Figure 4). We conducted a validation study using an independent data set to evaluate the top 2 findings at rs6733839 and rs1582763. The result is shown in Table 5, where both interaction findings were replicated using this independent data set, showing consistent directions of RERIs with improved significance when data were pooled for discovery and validation.

DISCUSSION

In this study, we developed a robust LRT for detecting G × E interaction based on the trend effect of genotype under an additive risk model that incorporates the G-E independence assumption. We used a constrained likelihood approach to impose 2 sets of constraints, the linear trend

Downloaded from https://academic.oup.com/aje/article/190/1/129/5900107 by University of Southern California user on 10 June 2021

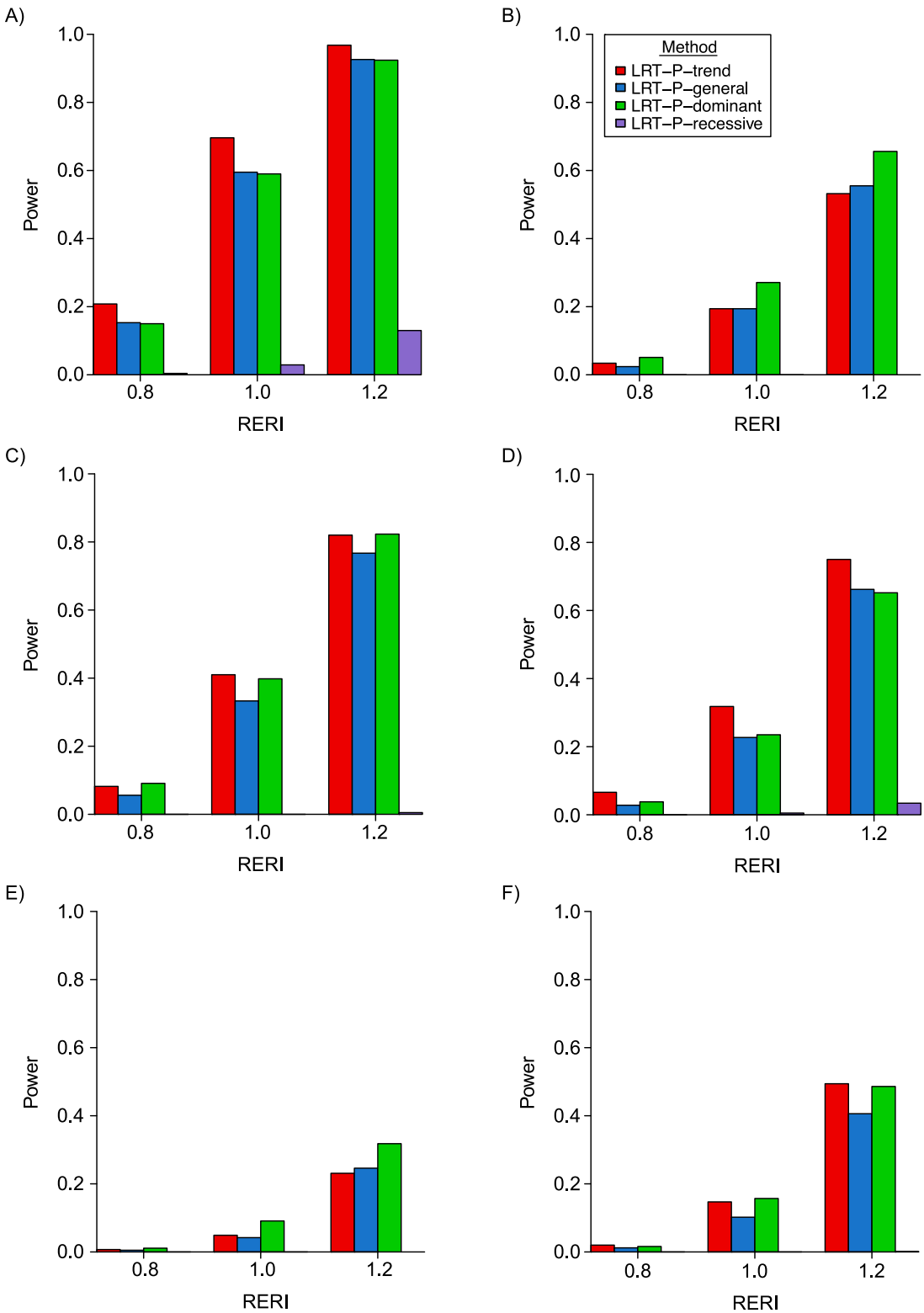

Figure 1. The results for power simulation of additive interaction tests based on a prospective likelihood for data generated under trend model and marginal odds ratio (MOR)(E) = 2.5 (A), dominant model and MOR(E) = 2.5 (B), general model and MOR(E) = 2.5 (C), trend model and MOR(E) = 3 (D), dominant model and MOR(E) = 3 (E), general model and MOR(E) = 3 (F). Significance level of $\alpha = 1 \times 10^{-7}$ was used. 1,000 replicated data sets were simulated for 5,000 cases and 5,000 controls. For each set of simulation, we applied the following 4 additive interaction tests based on a prospective likelihood: the likelihood ratio test (LRT) under the trend effect of genotype (prospective LRT, LRT-P-trend), a general model (LRT-P-general), a dominant model (LRT-P-dominant), and a recessive model (LRT-P-recessive). RERI, relative excess risk due to interaction.

Downloaded from https://academic.oup.com/aje/article/190/1/129/5900107 by University of Southern California user on 10 June 2021

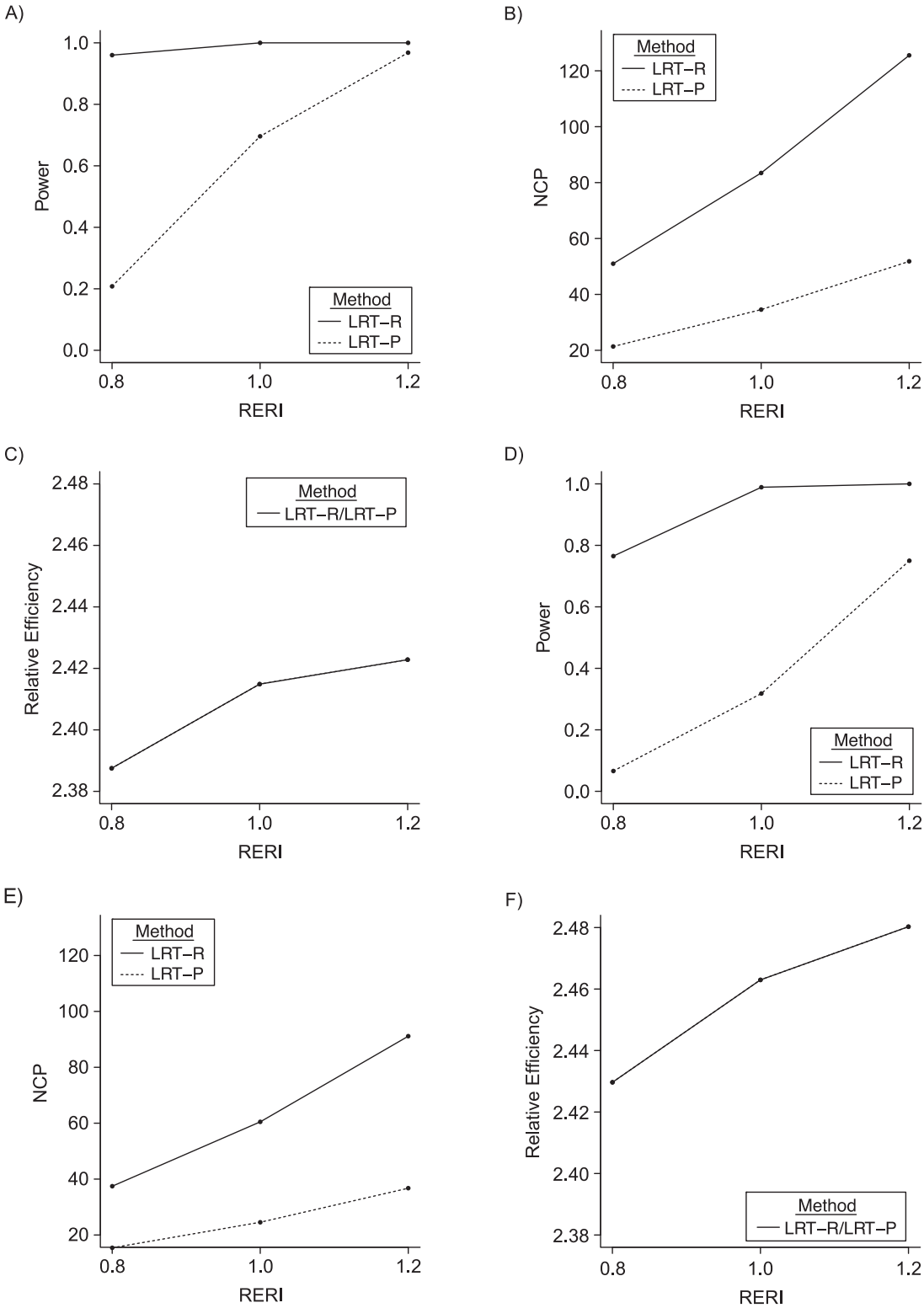

Figure 2. Comparison of the power of the trend effect–based additive interaction tests for the retrospective likelihood ratio test (LRT-R) versus prospective likelihood ratio test (LRT-P) for data generated with marginal odds ratio (E) (MOR(E)) = 2.5 (A) and MOR(E) = 3 (D). The noncentrality parameter (NCP) for each LRT, for MOR(E) = 2.5 (B) and MOR(E) = 3 (E), was estimated to compare the performances of the tests regardless of significance levels. The relative efficiency of LRT-R with regard to LRT-P, for MOR(E) = 2.5 (C) and MOR(E) = 3 (F), was estimated by taking the ratio of the NCP of LRT-R to the NCP of LRT-P. RERI, relative excess risk due to interaction.

Downloaded from https://academic.oup.com/aje/article/190/1/129/5900107 by University of Southern California user on 10 June 2021

Table 4. *P* Value of Lung Cancer Data Analysis[a] Using a Set of Additive Interaction Tests Based on the Trend Effect of the Genotype, Dominant Model, and General Model

SNP	Chromosomal Region	LRT-R[b]			LRT-P		
		Trend	Dominant	General	Trend[c]	Dominant	General
rs8034191	15q25.1[d]	N/A	N/A	N/A	5.55×10^{-16}[e]	1.13×10^{-14}[e]	7.88×10^{-15}[e]
rs8042374	15q25.1[d]	N/A	N/A	N/A	2.77×10^{-12}[e]	1.37×10^{-11}[e]	1.49×10^{-11}[e]
rs3117582	6p21.33	4.07×10^{-3}	1.28×10^{-2}	3.24×10^{-3}	3.08×10^{-3}	1.46×10^{-2}	1.23×10^{-3}
rs31489	5p15.33	4.11×10^{-4}[e]	1.18×10^{-2}	1.67×10^{-3}	1.84×10^{-2}	6.39×10^{-2}	5.73×10^{-2}
rs2395185	6p21.32	1.81×10^{-2}	3.10×10^{-2}	2.88×10^{-2}	4.02×10^{-2}	3.75×10^{-2}	1.10×10^{-1}
rs4324798	6p22.1	2.76×10^{-1}	4.76×10^{-1}	2.63×10^{-2}	6.12×10^{-2}	1.39×10^{-1}	3.98×10^{-2}
rs4975616	5p15.33	3.77×10^{-3}	5.76×10^{-3}	1.76×10^{-2}	7.11×10^{-2}	4.02×10^{-2}	1.19×10^{-1}
rs401681	5p15.33	1.94×10^{-3}	4.98×10^{-2}	3.11×10^{-3}	7.16×10^{-2}	1.89×10^{-1}	1.77×10^{-1}
rs2736100	5p15.33	5.17×10^{-2}	8.98×10^{-2}	1.27×10^{-1}	2.56×10^{-1}	5.22×10^{-1}	4.79×10^{-1}
rs7216064	17q24.2	5.46×10^{-1}	7.12×10^{-1}	8.10×10^{-1}	3.16×10^{-1}	4.82×10^{-1}	4.27×10^{-1}
rs3817963	6p21.32	1.19×10^{-1}	1.98×10^{-1}	2.94×10^{-1}	3.65×10^{-1}	5.16×10^{-1}	5.86×10^{-1}
rs9387478	6q22.1	9.68×10^{-1}	5.29×10^{-1}	6.10×10^{-1}	4.29×10^{-1}	4.69×10^{-1}	7.20×10^{-1}
rs10937405	3q28	9.84×10^{-1}	8.74×10^{-1}	9.64×10^{-1}	8.92×10^{-1}	6.10×10^{-1}	5.06×10^{-1}
rs753955	13q12.12	7.68×10^{-1}	9.06×10^{-1}	9.58×10^{-1}	9.64×10^{-1}	8.36×10^{-1}	8.82×10^{-1}

Abbreviations: LRT-P, prospective likelihood ratio test; LRT-R, retrospective likelihood ratio test; N/A, not applicable; SNP, single nucleotide polymorphism.

[a] Using genome-wide association study data from the National Cancer Institute (Landi et al. (27)).

[b] Retrospective likelihood analysis assumes G-E independence.

[c] The rows are sorted by the *P* value of LRT-P.

[d] The SNPs in 15q25.1 are known to be associated with smoking (E) and hence not tested using the retrospective likelihood due to violation of G-E independence.

[e] Statistically significant *P* values; $\alpha = 0.0006$ ($= 0.05/(14 \times 6)$) was applied for statistical significance.

effect of genotype and the additive joint effect of gene and environment based on a saturated logit model. Our simulation study demonstrated that the proposed test is robust across different underlying genetic models, showing increased power compared with alternative methods based on the dominant, recessive, or general models. The proposed trend-based interaction test using the retrospective likelihood showed approximately 2.5-fold increased efficiency compared with the method based on the standard prospective likelihood when the G-E independence assumption holds.

Application of the proposed method to the GWAS data for LOAD yielded 2 significant interactions at genome-wide significance between rs1582763 and *APOE* ε4 and rs6733839 and *APOE* ε4, which were replicated using independent data. Notably, both of these SNPs were previously identified to be associated with LOAD risk using GWAS data (48, 49). rs1582763 is located in the membrane-spanning 4-domains subfamily A (*MS4A*) gene cluster (closest to *MS4A* member 4A (*MS4A4A*), on chromosome 11q12.2, Online Mendelian Inheritance in Man ∗606547), and rs6733839 is located near/in the bridging integrator 1 gene (*BIN1*, on chromosome 2q14.3, Online Mendelian Inheritance in Man ∗601248). The A allele of rs1582763 was found to be associated with a decreased risk of LOAD ($P = 4.72 \times 10^{-9}$) using proxy-phenotype analysis of GWAS with subjects with parental LOAD status (48), and further *MS4A4A* was implicated in AD family history-based GWAS (50), LOAD GWAS (45), and sporadic AD GWAS (51). rs6733839, located near *BIN1*, was associated with LOAD risk (49–51) (odds ratio = 1.22; $P = 6.9 \times 10^{-44}$ in Lambert et al. (49)), and *BIN1* was further implicated in LOAD GWAS (45, 52) and in sporadic AD GWAS (51). While a recent study based on 53,711 subjects (46) showed that both *MS4A* and *BIN1* had potentially differential associations with LOAD risk among *APOE* ε4 carriers versus noncarriers, none attained significant SNP × *APOE* ε4 (multiplicative) interactions ($P = 0.27$ and $P = 0.87$ for *MS4A* and *BIN1* respectively). On the other hand, our analysis detected additive interactions at both of these loci at genome-wide significance. While we also examined multiplicative interactions (data not shown) for the top 30 SNPs obtained from SNP × *APOE* ε4 analysis, none of them, including rs1582763 and rs6733839, showed statistical significance using $\alpha = 5 \times 10^{-8}$ or $\alpha = 1 \times 10^{-5}$ (using both retrospective and prospective likelihoods). This implies that either a multiplicative risk model holds well overall for explaining the joint effect of each SNP and *APOE* ε4 for LOAD, or a multiplicative test is underpowered to detect true SNP × *APOE* ε4 interactions due to potentially smaller interaction effect sizes (versus RERIs that can be detected using additive interaction). We further note that besides the top 2 SNPs, 2 other SNPs,

Downloaded from https://academic.oup.com/aje/article/190/1/129/5900107 by University of Southern California user on 10 June 2021

Table 5. Odds Ratio for Each Genotype and Apolipoprotein E ε4 Allele Status Under the Trend Effect of Genotype for the Top 2 Single-Nucleotide Polymorphisms That Exceed the Genome-Wide Association Study Significance Level ($P < 5 \times 10^{-8}$) for Alzheimer Disease Data Under the Retrospective Likelihood[a]

Data Source and SNP	No.	Chromosome	Gene	Genotype	*APOE* ε4 Carrier		RERI	95% CI	*P* for Interaction
					No	Yes			
Discovery data	16,474								
rs6733839		2	*BIN1*	CC	1.00	5.04			
				CT	1.26	6.41	1.11	(0.788, 1.422)	5.56×10^{-14}
				TT	1.53	7.78			
rs1582763		11	*MS4A4A*	GG	1.00	5.19			
				GA	0.88	4.43	−0.64	(−0.852, −0.426)	5.32×10^{-9}
				AA	0.76	3.68			
Validation data	3,857								
rs6733839		2	*BIN1*	CC	1.00	3.39			
				CT	1.19	4.02	0.44	(0.011, 0.87)	0.036
				TT	1.38	4.65			
rs1582763		11	*MS4A4A*	GG	1.00	3.25			
				GA	0.78	2.67	−0.36	(−0.637, −0.078)	0.011
				AA	0.55	2.08			
Pooled data	20,331								
rs6733839		2	*BIN1*	CC	1.00	4.56			
				CT	1.25	5.74	0.93	(0.676,1.187)	1.17×10^{-14}
				TT	1.50	6.92			
rs1582763		11	*MS4A4A*	GG	1.00	4.64			
				GA	0.86	3.93	−0.57	(−0.743, −0.401)	$5.97E \times 10^{-11}$
				AA	0.72	3.22			

Note: Abbreviations: APOE, apolipoprotein E; BIN1, bridging integrator 1; CI, confidence interval; MS4A4A, membrane-spanning 4-domains subfamily A member 4A; RERI, relative excess risk due to interaction; SNP, single nucleotide polymorphism.

[a] The results under the prospective likelihood are shown in Web Table 5. Data made available by National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site and similar LOAD repositories (Web Appendix 1; Web Tables 1–2) (45).

Downloaded from https://academic.oup.com/aje/article/190/1/129/5900107 by University of Southern California user on 10 June 2021

rs1752684 and rs12444073, retained statistical significance using a suggestive threshold ($\alpha = 1 \times 10^{-5}$) (Web Table 4). Specifically, zinc finger protein 423 (*ZNF423*, where rs12444073 is located) has recently been shown to be associated with LOAD risk (gene-level $P = 2.1 \times 10^{-6}$) (53).

To the best of our knowledge, our study presents the first method that incorporates the trend effect of genotype for testing additive interaction using both retrospective and prospective likelihoods. Previously, an LRT has been proposed for testing additive interaction that assumes either binary genetic data (dominant or recessive models) or general model, which has been applied to various complex diseases including bladder cancer and breast cancer (9–11). However, the major limitation of this method was strong assumptions from genetic models. Application of the proposed trend effect–based method that overcomes this limitation has led to reproducible interaction findings for LOAD. Last, our new method is implemented in the freely available R package, CGEN (24), which can facilitate its wide application among researchers in molecular epidemiology.

Despite these strengths, our study has limitations. While the proposed test based on the retrospective likelihood is known to increase power when the assumption of G-E (or G-G) independence holds, it should be used with caution because violation of the assumption can produce a bias (54). To handle this issue, an empirical Bayes-type shrinkage estimator was previously developed for examining multiplicative interaction (55). This method employed a weighted average of the retrospective and prospective likelihood-based estimators for multiplicative interaction, yielding an acceptable trade-off between bias and efficiency. A recent work proposed a similar approach for additive interaction, assuming a dominant or recessive model (14). Currently, an extension of our proposed trend effect–based method is under way to incorporate an empirical Bayes-type shrinkage estimator.

To conclude, we have developed a test for detecting additive G-E interaction based on the trend effect of genotype and extended it to exploit the independence between gene and environment. Our simulation study shows that the proposed method is robust under varying genetic models and improves power when the assumption on G-E independence is held. Future work is needed to relax this strong assumption to be more data-adaptive.

ACKNOWLEDGMENTS

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

S.S.H. and N.C. made an equal contribution to this work.

This work was funded by the National Cancer Institute (grant 1R37CA226081 to S.S.H.) and the National Institutes of Health (grants AG060747 and AG047366 to M.D.G.).

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of Alzheimer's Disease Neuroimaging Initiative and/or provided data but did not participate in analysis or writing of this report. A complete listing of Alzheimer's Disease Neuroimaging Initiative investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. Biological samples used in this study were stored at principal investigators' institutions and at the National Cell Repository for Alzheimer's Disease at Indiana University, which receives government support under a cooperative agreement grant (U24 AG21886) awarded by the National Institute on Aging). 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;

Downloaded from https://academic.oup.com/aje/article/190/1/129/5900107 by University of Southern California user on 10 June 2021

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 NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, U01AG46152, the Illinois Department of Public Health, and the Translational Genomics Research Institute. 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.

Conflict of interest: none declared.

REFERENCES

1. Thomas D. Gene–environment-wide association studies: emerging approaches. *Nat Rev Genet*. 2010;11(4):259–272.
2. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. *Am J Epidemiol*. 1980;112(4):467–470.
3. Garcia-Closas M, Gunsoy NB, Chatterjee N. Combined associations of genetic and environmental risk factors: implications for prevention of breast cancer. *J Natl Cancer Inst*. 2014;106(11):dju305.
4. Lund E. Comparison of additive and multiplicative models for reproductive risk factors and post-menopausal breast cancer. *Stat Med*. 1995;14(3):267–274.
5. Thompson WD. Effect modification and the limits of biological inference from epidemiologic data. *J Clin Epidemiol*. 1991;44(3):221–232.
6. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. *Stat Med*. 1983;2(2):243–251.
7. Walter S, Holford T. Additive, multiplicative, and other models for disease risks. *Am J Epidemiol*. 1978;108(5):341–346.
8. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet*. 2002;11(20):2463–2468.
9. Figueroa JD, Han SS, Garcia-Closas M, et al. Genome-wide interaction study of smoking and bladder cancer risk. *Carcinogenesis*. 2014;35(8):1737–1744.
10. Fu Y-P, Kohaar I, Rothman N, et al. Common genetic variants in the PSCA gene influence gene expression and bladder cancer risk. *Proc Natl Acad Sci*. 2012;109(13):4974–4979.
11. Joshi AD, Lindström S, Hüsing A, et al. Additive interactions between susceptibility single-nucleotide polymorphisms identified in genome-wide association studies and breast cancer risk factors in the Breast and Prostate Cancer Cohort Consortium. *Am J Epidemiol*. 2014;180(10):1018–1027.
12. Han SS, Chatterjee N. Review of statistical methods for gene-environment interaction analysis. *Curr Epidemiol Rep*. 2018;5(1):39–45.
13. 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.
14. 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.
15. Tchetgen Tchetgen EJ, Shi X, Wong BH, 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.
16. 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.
17. Chu H, Nie L, Cole SR. Estimating the relative excess risk due to interaction: a Bayesian approach. *Epidemiology*. 2011;22(2):242–248.
18. Nie L, Chu H, Li F, et al. Relative excess risk due to interaction: resampling-based confidence intervals. *Epidemiology*. 2010;21(4):552–556.
19. 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.
20. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447(7145):661–678.
21. Freidlin B, Zheng G, Li Z, et al. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered*. 2002;53(3):146–152.
22. Timpson NJ, Greenwood CM, Soranzo N, et al. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat Rev Genet*. 2018;19(2):110–124.
23. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*. 2014;42(D1):D1001–D1006.
24. Bhattacharjee S, Chatterjee C, Han S, et al. CGEN: an R package for analysis of case-control studies in genetic epidemiology. [R package version 3180]. https://bioconductor.statistik.tu-dortmund.de/packages/3.8/bioc/html/CGEN.html. Accessed September 4, 2020.
25. Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*. 2005;92(2):399–418.
26. 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.

Downloaded from https://academic.oup.com/aje/article/190/1/129/5900107 by University of Southern California user on 10 June 2021

27. 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.
28. 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.
29. 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.
30. McKay JD, Hung RJ, Gaborieau V, et al. Lung cancer susceptibility locus at 5p15.33. *Nat Genet*. 2008;40(12): 1404–1406.
31. 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.
32. Hu Z, Wu C, Shi Y, et al. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nat Genet*. 2011;43(8): 792–796.
33. Jin G, Ma H, Wu C, et al. Genetic variants at 6p21.1 and 7p15.3 are associated with risk of multiple cancers in Han Chinese. *Am J Hum Gen*. 2012;91(5):928–934.
34. Lan Q, Hsiung CA, Matsuo K, et al. Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nat Genet*. 2012; 44(12):1330–1335.
35. David S, Hamidovic A, Chen G, et al. Genome-wide meta-analyses of smoking behaviors in African Americans. *Transl Psychiatry*. 2012;2(5):e119.
36. 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*. [published online ahead of print January 7, 2019]. (doi: 10.1038/s41380-018-0313-0).
37. The Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet*. 2010;42(5):441–447.
38. 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.
39. Liu M, Jiang Y, Wedow R, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet*. 2019;51(2): 237–244.
40. 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.
41. 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.
42. Belloy ME, Napolioni V, Greicius MD. A quarter century of APOE and Alzheimer's disease: progress to date and the path forward. *Neuron*. 2019;101(5):820–838.
43. 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.
44. Altmann A, Tian L, Henderson VW, et al. Sex modifies the APOE-related risk of developing Alzheimer disease. *Ann Neurol*. 2014;75(4):563–573.
45. Naj AC, Jun G, Beecham GW, et al. Common variants at *MS4A4/MS4A6E*, *CD2AP*, *CD33* and *EPHA1* are associated with late-onset Alzheimer's disease. *Nat Genet*. 2011;43(5): 436–441.
46. 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.
47. 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.
48. Jansen IE, Savage JE, Watanabe K, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet*. 2019;51(3): 404–413.
49. Lambert J-C, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet*. 2013; 45(12):1452–1458.
50. Marioni RE, Harris SE, Zhang Q, et al. GWAS on family history of Alzheimer's disease. *Transl Psychiatry*. 2018; 8(1):99.
51. Antúnez C, Boada M, González-Pérez A, et al. The membrane-spanning 4-domains, subfamily A (MS4A) gene cluster contains a common variant associated with Alzheimer's disease. *Genome Med*. 2011;3(5):33.
52. Hu X, Pickering E, Liu YC, et al. Meta-analysis for genome-wide association study identifies multiple variants at the *BIN1* locus associated with late-onset Alzheimer's disease. *PloS One*. 2011;6(2):e16616.
53. Baker E, Sims R, Leonenko G, et al. Gene-based analysis in HRC imputed genome wide association data identifies three novel genes for Alzheimer's disease. *PloS One*. 2019; 14(7):e0218111.
54. 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.
55. 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.

Downloaded from https://academic.oup.com/aje/article/190/1/129/5900107 by University of Southern California user on 10 June 2021