Title
Robust Tests for Additive Gene-Environment Interaction in Case-Control Studies Using Gene-Environment Independence.

Journal
American journal of epidemiology, 187(2)

ISSN
0002-9262

Authors
Liu, Gang
Mukherjee, Bhramar
Lee, Seunggeun
et al.

Publication Date
2018-02-01

DOI
10.1093/aje/kwx243

American Journal of Epidemiology

Vol. 187, No. 2
DOI: 10.1093/aje/kwx243
Advance Access publication: June 14, 2017

Practice of Epidemiology

Robust Tests for Additive Gene-Environment Interaction in Case-Control Studies Using Gene-Environment Independence

Gang Liu, Bhramar Mukherjee*, Seunggeun Lee, Alice W. Lee, Anna H. Wu, Elisa V. Bandera, Allan Jensen, Mary Anne Rossing, Kirsten B. Moysich, Jenny Chang-Claude, Jennifer A. Doherty, Aleksandra Gentry-Maharaj, Lambertus Kiemeney, Simon A. Gayther, Francesmary Modugno, Leon Massuger, Ellen L. Goode, Brooke L. Fridley, Kathryn L. Terry, Daniel W. Cramer, Susan J. Ramus, Hoda Anton-Culver, Argyrios Ziogas, Jonathan P. Tyrer, Joellen M. Schildkraut, Susanne K. Kjaer, Penelope M. Webb, Roberta B. Ness, Usha Menon, Andrew Berchuck, Paul D. Pharoah, Harvey Risch, and Celeste Leigh Pearce, for the Ovarian Cancer Association Consortium

* Correspondence to Dr. Bhramar Mukherjee, Department of Biostatistics and Epidemiology, School of Public Health, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109 (email: bhramar@umich.edu).

Initially submitted May 20, 2016; accepted for publication June 2, 2017.

There have been recent proposals advocating the use of additive gene-environment interaction instead of the widely used multiplicative scale, as a more relevant public health measure. Using gene-environment independence enhances statistical power for testing multiplicative interaction in case-control studies. However, under departure from this assumption, substantial bias in the estimates and inflated type I error in the corresponding tests can occur. In this paper, we extend the empirical Bayes (EB) approach previously developed for multiplicative interaction, which trades off between bias and efficiency in a data-adaptive way, to the additive scale. An EB estimator of the relative excess risk due to interaction is derived, and the corresponding Wald test is proposed with a general regression setting under a retrospective likelihood framework. We study the impact of gene-environment association on the resultant test with case-control data. Our simulation studies suggest that the EB approach uses the gene-environment independence assumption in a data-adaptive way and provides a gain in power compared with the standard logistic regression analysis and better control of type I error when compared with the analysis assuming gene-environment independence. We illustrate the methods with data from the Ovarian Cancer Association Consortium.

bias-variance tradeoff; effect modification; empirical Bayes estimation; genetic risk score; relative excess risk; shrinkage

Abbreviations: CI, confidence interval; CML, constrained maximum likelihood; EB, empirical Bayes; GRS, genetic risk score; LRT, likelihood ratio test; MLE, maximum likelihood estimate; OCP, oral contraceptive pill; RERI, relative excess risk due to interaction; SNP, single nucleotide polymorphism; UML, unconstrained maximum likelihood; WGRS, weighted genetic risk score.

There has been increasing interest in searching for gene × environment ($G \times E$) interaction in the post–genome-wide association studies era, with limited success (1–5). A number of methods have been proposed for efficient searching for $G \times E$ effects that use the gene-environment independence assumption (2, 6–10). Almost all of these studies have focused on testing/estimation of multiplicative interaction, perhaps due to the fact that standard logistic regression is the most commonly used tool for analyzing case-control data (11–13). However, it has been suggested in the literature that additive interaction is a more relevant public health measure (3, 14, 15). If an environmental exposure, say E, can potentially be modified via an intervention, the additive $G \times E$ interaction measure can quantify the differences in the number of cases prevented if the intervention was offered in a prioritized way, across strata defined by genetic risk. This characterization helps with policy questions when

there is limited access to an intervention. Moreover, the additive measure of interaction corresponds more closely to the notion of mechanistic or causal measures of interaction (16, 17).

Although this is not commonly recognized, it is possible to *test* for additive interaction in a logistic regression model using case-control data. While a direct estimate of additive interaction on a risk difference scale cannot be obtained from case-control data, an alternative parameter, the relative excess risk due to interaction (RERI), can be represented in terms of relative risks. Assuming that the disease is rare, relative risks can be approximated by corresponding odds ratios, and thus the RERI can be viewed as a function of both main effects and multiplicative interaction parameters in a logistic regression model. The standard delta theorem can be applied to provide asymptotic variance, and subsequently a Wald test for the null hypothesis RERI = 0 can be conducted (18–20). The fact that RERI = 0 if and only if the additive null hypothesis holds provides us with a way to test for interaction on the additive scale by testing H_0: RERI = 0. More recently, Han et al. (21) developed a likelihood ratio test (LRT) for H_0: RERI = 0, applying the retrospective likelihood framework proposed by Chatterjee and Carroll (22) that permits incorporation of the *G-E* independence assumption and leads to a more powerful test than the previously proposed Wald test, in modest sample sizes, for both the unconstrained and constrained maximum likelihood methods. However, it is not clear how to extend the LRT in an empirical Bayes-type adaptive framework, and thus we proceeded with combining estimates of RERI instead of deriving a combination LRT.

In this paper, we first consider the binary *G*, *E* scenario to illustrate our method for testing additive interaction in case-control studies. We provide closed-form expressions of the maximum likelihood estimates (MLEs) and the Wald test of the RERI parameter while assuming gene-environment independence (constrained MLE) and not assuming gene-environment independence (unconstrained MLE). We then extend the empirical Bayes-type shrinkage approach for multiplicative $G \times E$ interaction proposed by Mukherjee and Chatterjee (6) to estimate RERI and test for additive interaction. An adaptively weighted estimator of RERI that combines the constrained and unconstrained estimators is proposed to trade off between bias and efficiency. Finally, we extend the method to handle a completely general regression setting using the retrospective profile likelihood-based framework in Chatterjee and Carroll (22). We conduct a simulation study to compare the performance of various tests and illustrate our method by applying it to study of the interaction between use of oral contraceptive pills (OCPs) and previously identified genetic factors in a large consortium of case-control studies of ovarian cancer.

METHODS

We first consider a simple setup of an unmatched case-control study with a dichotomous genetic factor *G* and a dichotomous environmental exposure *E*. Let $E = 1$ ($E = 0$) denote an exposed (unexposed) individual and $G = 1$ ($G = 0$) denote whether an individual is a carrier (noncarrier) of the susceptible genetic marker. Let *D* denote disease status, where $D = 1$ ($D = 0$) stands for an affected (unaffected) individual. Let N_0 and N_1 be the numbers of selected controls and cases, respectively. The data can be represented in the form of a 2×4 table as displayed in Web Appendix 1 (available at https://academic.oup.com/aje).

Let $\boldsymbol{r}_0 = (r_{01}, r_{02}, r_{03}, r_{04})$ and $\boldsymbol{r}_1 = (r_{11}, r_{12}, r_{13}, r_{14})$ denote the vectors of observed cell frequencies in the controls and the cases, respectively. Let $r_G = r_{03} + r_{04}$ denote the frequency of $G = 1$ and $r_E = r_{02} + r_{04}$ denote the frequency of $E = 1$ among controls. Let $\boldsymbol{p}_0 = (p_{01}, p_{02}, p_{03}, p_{04})$ and $\boldsymbol{p}_1 = (p_{11}, p_{12}, p_{13}, p_{14})$ denote the true population parameters of the cell probabilities corresponding to a particular *G-E* configuration in the underlying control and case populations, respectively. Let $p_G = p_{03} + p_{04}$ denote the marginal prevalence of $G = 1$ among controls and $p_E = p_{02} + p_{04}$ denote the marginal prevalence of $E = 1$ among controls. The observed vectors of the cell counts can be viewed as random draws from 2 independent multinomial distributions in controls and cases, respectively, namely $\boldsymbol{r}_0 \sim$ multinomial $(N_0, \boldsymbol{p}_0)$ and $\boldsymbol{r}_1 \sim$ multinomial $(N_1, \boldsymbol{p}_1)$.

Let us introduce the following notation for the key parameters of interest. Let

$$\mathrm{OR}_E = \frac{P(D=1|E=1,G=0)}{P(D=0|E=1,G=0)} \Big/ \frac{P(D=1|E=0,G=0)}{P(D=0|E=0,G=0)} = p_{01}p_{12}/p_{02}p_{11}$$

denote the odds ratio (OR) associated with *E* for nonsusceptible individuals ($G = 0$),

$$\mathrm{OR}_G = \frac{P(D=1|G=1,E=0)}{P(D=0|G=1,E=0)} \Big/ \frac{P(D=1|G=0,E=0)}{P(D=0|G=0,E=0)} = p_{01}p_{13}/p_{03}p_{11}$$

denote the odds ratio associated with *G* for unexposed individuals ($E = 0$), and

$$\mathrm{OR}_{GE} = \frac{P(D=1|E=1,G=1)}{P(D=0|E=1,G=1)} \Big/ \frac{P(D=1|E=0,G=0)}{P(D=0|E=0,G=0)} = p_{01}p_{14}/p_{04}p_{11}$$

denote the joint odds ratio associated with the subgroup $G = 1$ and $E = 1$ compared with the reference group of $G = 0$ and $E = 0$. The multiplicative interaction parameter ψ is defined as

$$\psi = \frac{\mathrm{OR}_{GE}}{\mathrm{OR}_G \mathrm{OR}_E} = \frac{p_{02}p_{03}p_{11}p_{14}}{p_{01}p_{04}p_{12}p_{13}} = \frac{(p_{11}p_{14}/p_{12}p_{13})}{\exp(\theta_{GE})},$$

where

$$\theta_{GE} = \log \frac{p_{01}p_{04}}{p_{02}p_{03}}.$$

The parameter θ_{GE} represents the log odds ratio between *G* and *E* among the controls, characterizing the gene-environment association. On the additive scale, the measure of interaction is defined as

$$\begin{aligned} p_{\text{additive}} = &[P(D=1|E=1, G=1) \\ &- P(D=1|E=0, G=0)] \\ &- [P(D=1|E=1, G=0) \\ &- P(D=1|E=0, G=0)] \\ &- [P(D=1|E=0, G=1) \\ &- P(D=1|E=0, G=0)] \\ = &P(D=1|E=1, G=1) \\ &- P(D=1|E=1, G=0) \\ &- P(D=1|E=0, G=1) \\ &+ P(D=1|E=0, G=0). \end{aligned} \quad (1)$$

Dividing equation 1 throughout by $P(D=1|E=0, G=0)$, we obtain a new measure, the RERI:

$$\text{RERI}_{\text{RR}} = \text{RR}_{GE} - \text{RR}_G - \text{RR}_E + 1. \quad (2)$$

When the disease is rare, the odds ratio approximates the relative risk (RR). Hence, we have

$$\text{RERI}_{\text{OR}} \approx \text{OR}_{GE} - \text{OR}_G - \text{OR}_E + 1. \quad (3)$$

Note that by equations 1 and 3, testing H_0: $p_{\text{additive}} = 0$ is equivalent to testing H_0: $\text{RERI}_{\text{RR}} = 0$, which is typically translated into H_0: $\text{RERI}_{\text{OR}} = 0$ in a case-control study, as described by VanderWeele (23). After defining the above relevant parameters of interest, we use the definition of RERI in equation 3 in terms of odds ratios to proceed with inference under case-control sampling, assuming that the disease is rare for all configurations of G and E.

Unconstrained maximum likelihood estimation

The unconstrained maximum likelihood (UML) estimate for all odds ratio parameters mentioned above is obtained by simply substituting p_{dj} with its MLE, $\hat{p}_{dj} = r_{dj}/N_d$, implying

$$\hat{\psi}_{\text{UML}} = \frac{\widehat{\text{OR}}_{GE}}{\widehat{\text{OR}}_G \widehat{\text{OR}}_E} = \frac{r_{02}r_{03}r_{11}r_{14}}{r_{01}r_{04}r_{12}r_{13}}$$

and

$$\hat{\sigma}^2_{\text{UML}} = \text{Var}(\log(\hat{\psi}_{\text{UML}})) = \sum_{d=0}^{1}\sum_{j=1}^{4}\frac{1}{r_{dj}}.$$

The G-E association log odds ratio in controls can also be estimated as $\hat{\theta}_{GE} = \log(r_{01}r_{04}/r_{02}r_{03})$.

The UML estimate of RERI can be easily obtained by plugging the corresponding estimated odds ratios in an unconstrained model into equation 3, and by the invariance property of MLE, it serves as a consistent and asymptotically unbiased estimate of RERI regardless of the gene-environment independence assumption.

$$\widehat{\text{RERI}}_{\text{UML}} = \frac{r_{01}r_{14}}{r_{11}r_{04}} - \frac{r_{01}r_{13}}{r_{11}r_{03}} - \frac{r_{01}r_{12}}{r_{11}r_{02}} + 1 \quad (4)$$

Note that $\boldsymbol{r}_0$ and $\boldsymbol{r}_1$ are realizations from 2 independent multinomial distributions, and we can employ the delta method (Web Appendix 2) to obtain the asymptotic variance of $\widehat{\text{RERI}}_{\text{UML}}$, which is the same as noted in references 17–19. The Wald test for interaction is based on the standardized Z statistic

$$Z_{\text{UML}} = \widehat{\text{RERI}}_{\text{UML}}/\sqrt{\widehat{\text{Var}}(\widehat{\text{RERI}}_{\text{UML}})},$$

which follows an $N(0, 1)$ distribution under the null RERI = 0.

Constrained maximum likelihood estimation

Under G-E independence among controls (i.e., $\theta_{GE} = 0$) and the rare disease assumption, Zhang et al. (24) proposed the constrained maximum likelihood (CML) estimates for $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ as follows:

$$\hat{p}_{01} = \frac{(r_{01}+r_{03})(r_{01}+r_{02})}{N_0^2},\ \hat{p}_{02} = \frac{(r_{01}+r_{02})(r_{02}+r_{04})}{N_0^2},$$
$$\hat{p}_{03} = \frac{(r_{01}+r_{03})(r_{03}+r_{04})}{N_0^2},\ \hat{p}_{04} = \frac{(r_{02}+r_{04})(r_{03}+r_{04})}{N_0^2},$$

and

$$\hat{p}_{1j} = \frac{r_{1j}}{N_1},\quad j = 1, 2, 3, 4.$$

We obtain the corresponding odds ratio estimates by substituting p_{dj} with its constrained MLE under G-E independence:

$$\widehat{\text{OR}}_E = \frac{r_{12}(r_{01}+r_{03})}{r_{11}(r_{02}+r_{04})},\ \widehat{\text{OR}}_G = \frac{r_{13}(r_{01}+r_{02})}{r_{11}(r_{03}+r_{04})},$$
$$\widehat{\text{OR}}_{GE} = \frac{r_{14}(r_{01}+r_{02})(r_{01}+r_{03})}{r_{11}(r_{02}+r_{04})(r_{03}+r_{04})},$$

and

$$\hat{\psi}_{\text{CML}} = \frac{r_{11}r_{14}}{r_{12}r_{13}},\ \hat{\sigma}^2_{\text{CML}} = \text{Var}(\log(\hat{\psi}_{\text{CML}})) = \sum_{j=1}^{4}\frac{1}{r_{1j}}.$$

Note that the estimated multiplicative interaction parameter $\hat{\psi}$ is a function of only $\boldsymbol{r}_1$ and is identical to the case-only estimator. The CML estimate of RERI can be computed by plugging the estimated odds ratios under the constraint into equation 3. Formally, the CML estimator for RERI is given by

$$\begin{aligned} \widehat{\text{RERI}}_{\text{CML}} = &\frac{(r_{01}+r_{03})(r_{01}+r_{02})r_{14}}{(r_{02}+r_{04})(r_{03}+r_{04})r_{11}} - \frac{(r_{01}+r_{02})r_{13}}{(r_{03}+r_{04})r_{11}} \\ &- \frac{(r_{01}+r_{03})r_{12}}{(r_{02}+r_{04})r_{11}} + 1. \end{aligned} \quad (5)$$

Under the G-E independence assumption among controls, the CML estimator is consistent and asymptotically unbiased for the true RERI parameter. It is more precise than the UML estimator of RERI in equation 4 based on our simulations. The asymptotic variance of the CML estimator can also be approximated by means of the delta method, which is shown in Web Appendix 3. The Wald test for RERI in a constrained model again uses the standardized Z statistic

$$Z_{\mathrm{CML}} = \widehat{\mathrm{RERI}}_{\mathrm{CML}} / \sqrt{\widehat{\mathrm{Var}}(\widehat{\mathrm{RERI}}_{\mathrm{CML}})},$$

and the power of the test is slightly lower than that of the LRT for additive interaction in Han et al. (21), as will be illustrated through our simulations. Under violation of the *G-E* independence assumption, $\theta_{GE} \neq 0$, the CML estimate is asymptotically biased for the true RERI parameter and the tests are invalid.

Empirical Bayes estimation

Mukherjee and Chatterjee (6) proposed an empirical Bayes (EB) estimator of the multiplicative interaction which shrinks the UML and CML estimators in a data-adaptive way. It relaxes the *G-E* independence assumption and makes a tradeoff between bias and efficiency. Formally, the EB estimator of multiplicative interaction is given by

$$\log(\hat{\psi}_{\mathrm{EB}}) = \frac{\hat{\sigma}^2_{\mathrm{UML}}}{\hat{\theta}^2_{GE} + \hat{\sigma}^2_{\mathrm{UML}}} \log(\hat{\psi}_{\mathrm{CML}}) + \frac{\hat{\theta}^2_{GE}}{\hat{\theta}^2_{GE} + \hat{\sigma}^2_{\mathrm{UML}}} \log(\hat{\psi}_{\mathrm{UML}}), \quad (6)$$

where $\hat{\psi}_{\mathrm{CML}} = (r_{11}r_{14}/r_{12}r_{13})$, $\hat{\psi}_{\mathrm{UML}} = (r_{02}r_{03}r_{11}r_{14}/r_{01}r_{04}r_{12}r_{13})$, $\hat{\sigma}^2_{\mathrm{UML}} = \sum_{d=0}^{1}\sum_{j=1}^{4}(1/r_{dj})$, and $\hat{\theta}_{GE} = \log(r_{01}r_{04}/r_{02}r_{03})$.

We employ the same idea of adaptive weighting and propose the EB estimator for RERI as

$$\begin{aligned}\widehat{\mathrm{RERI}}_{\mathrm{EB}} &= \frac{(\widehat{\mathrm{RERI}}_{\mathrm{UML}} - \widehat{\mathrm{RERI}}_{\mathrm{CML}})^2}{\widehat{\mathrm{Var}}(\widehat{\mathrm{RERI}}_{\mathrm{UML}}) + (\widehat{\mathrm{RERI}}_{\mathrm{UML}} - \widehat{\mathrm{RERI}}_{\mathrm{CML}})^2}\widehat{\mathrm{RERI}}_{\mathrm{UML}} + \frac{\widehat{\mathrm{Var}}(\widehat{\mathrm{RERI}}_{\mathrm{UML}})}{\widehat{\mathrm{Var}}(\widehat{\mathrm{RERI}}_{\mathrm{UML}}) + (\widehat{\mathrm{RERI}}_{\mathrm{UML}} - \widehat{\mathrm{RERI}}_{\mathrm{CML}})^2}\widehat{\mathrm{RERI}}_{\mathrm{CML}} \\ &= \widehat{\mathrm{RERI}}_{\mathrm{UML}} + K(\widehat{\mathrm{RERI}}_{\mathrm{CML}} - \widehat{\mathrm{RERI}}_{\mathrm{UML}}), \end{aligned} \quad (7)$$

where $K = V(V + \hat{\kappa}\hat{\kappa}^T)^{-1}$ is a shrinkage factor of the same form as defined in Chen et al. (25) with $\hat{\kappa} = \widehat{\mathrm{RERI}}_{\mathrm{UML}} - \widehat{\mathrm{RERI}}_{\mathrm{CML}}$ and $V = \widehat{\mathrm{Var}}(\widehat{\mathrm{RERI}}_{\mathrm{UML}})$. To explain the intuitive rationale behind the estimator, observe that as $\widehat{\theta_{GE}} \to 0$—that is, as the data provide the evidence in favor of *G-E* independence ($\widehat{\mathrm{RERI}}_{\mathrm{UML}} - \widehat{\mathrm{RERI}}_{\mathrm{CML}} \to 0$)—the estimator puts more weight on the CML estimator to gain more efficiency; and as $\widehat{\theta_{GE}} \to \infty$—that is, as the *G-E* dependence becomes stronger in the control population and $\widehat{\mathrm{RERI}}_{\mathrm{UML}} - \widehat{\mathrm{RERI}}_{\mathrm{CML}}$ becomes larger—then the EB estimator puts more weight on the UML estimator to reduce bias. In large samples, the EB estimator converges to the UML estimate and thus is asymptotically unbiased for the true RERI parameter (6). The asymptotic variance of $\widehat{\mathrm{RERI}}_{\mathrm{EB}}$ is derived via the delta method (see Web Appendix 4), assuming $\widehat{\mathrm{Var}}(\widehat{\mathrm{RERI}}_{\mathrm{UML}})$ as a constant relative to the order of magnitude of the point estimates (6). We use the Wald test for the EB estimator based on the standardized *Z* statistic

$$Z_{\mathrm{EB}} = \widehat{\mathrm{RERI}}_{\mathrm{EB}} / \sqrt{\widehat{\mathrm{Var}}(\widehat{\mathrm{RERI}}_{\mathrm{EB}})}.$$

Remark 1. We also consider 2 other forms of adaptive weights. One is to modify the shrinkage factor *K* in equation 7 and let $\widehat{k^*} = \hat{\theta}_{GE}$ instead of $\widehat{\mathrm{RERI}}_{\mathrm{UML}} - \widehat{\mathrm{RERI}}_{\mathrm{CML}}$, namely,

$$\widehat{\mathrm{RERI}}_{\mathrm{EB1}} = \widehat{\mathrm{RERI}}_{\mathrm{UML}} + K^*(\widehat{\mathrm{RERI}}_{\mathrm{CML}} - \widehat{\mathrm{RERI}}_{\mathrm{UML}}),$$

where $K^* = V(V + \widehat{\kappa^*}\widehat{\kappa^*}^T)^{-1}$. The other is to plug in the EB estimates ($\widehat{\mathrm{OR}}_{\mathrm{EB}}$) obtained from using the retrospective likelihood framework in Mukherjee and Chatterjee (6), as implemented in R package CGEN (R Foundation for Statistical Computing, Vienna, Austria) (6, 22, 25), directly into equation 3, namely,

$$\widehat{\mathrm{RERI}}_{\mathrm{EB2}} = \widehat{\mathrm{OR}}_{GE} - \widehat{\mathrm{OR}}_{G} - \widehat{\mathrm{OR}}_{E} + 1,$$

where all estimated odds ratios are EB estimates proposed under the multiplicative model. The EB estimator we proposed in equation 7 demonstrates superior performance among the 3 choices, based on our simulation study.

Remark 2. As is shown in Chen et al. (25), the asymptotic theory for CML and consequently EB is nonregular under the independence assumption. The delta method does not technically apply for estimation of the asymptotic variance. Theoretically, the test statistic also fails to be asymptotically normal under *G-E* independence (25, 26). However, in practice, the estimated variance derived by the delta method approximates the empirical variance very well, as noted in the simulation studies (see Web Appendix 5, Web Tables 1 and 2, and Web Figures 1 and 2). Under *G-E* dependence, the EB estimate converges in large samples to the UML estimate and thus to the true RERI parameter, and standard likelihood asymptotics holds (6).

Profile likelihood framework for general regression setting

Consider the retrospective likelihood considered in the papers by Chatterjee and Carroll (22) and Mukherjee and Chatterjee (6) and as implemented in the R package CGEN:

$$\begin{aligned} &P(G, E, \mathbf{Z} \mid D) \\ &= \frac{P(D = 1 \mid G, E, \mathbf{Z})P(G \mid E, \mathbf{Z})P(E, \mathbf{Z})}{\sum_{G,E,\mathbf{Z}} P(D = 1 \mid G, E, \mathbf{Z})P(G \mid E, \mathbf{Z})P(E, \mathbf{Z})}. \end{aligned} \quad (8)$$

The 3 ingredients of the above retrospective likelihood are as follows.

1. The logistic regression disease risk model of interest with multiplicative $G \times E$ interaction parameter logit $P(D = 1 | G, E, \mathbf{Z}) = \beta_0 + \beta_G G + \beta_E E + \beta_{GE} G \times E + \boldsymbol{\beta}_Z^T \mathbf{Z}$, where $\mathbf{Z}$ denotes other covariates.
2. Logit $P(G | E, \mathbf{Z}) = \theta_0 + \theta_{GE} E + \boldsymbol{\theta}_{GZ}^T \mathbf{Z}$. While this is the gene model used for UML, allowing G-E dependence, in the CML method, $P(G | E, \mathbf{Z})$ reduces to $P(G | \mathbf{Z})$ under the assumption of G-E independence conditional on $\mathbf{Z}$, implying $\theta_{GE} \equiv 0$.
3. The distribution $P(E, \mathbf{Z})$ is allowed to be completely nonparametric. We then maximize the retrospective likelihood using existing routines in CGEN to obtain $\widehat{\beta}_{\text{UML}}$ and $\widehat{\beta}_{\text{CML}}$, the vector of all of the parameter estimates of the disease risk model in point 1 above, namely, $(\beta_0, \beta_G, \beta_E, \beta_{GE}, \boldsymbol{\beta}_Z)$.

When it comes to defining RERI with a general G and E variable adjusting for covariates $\mathbf{Z}$, particularly with case-control data, as described in VanderWeele (23), let us denote the RERI by $\text{RERI}_{\text{OR}}(E_0, E_1, G_0, G_1)$, by replacing risk ratios with corresponding odds ratios in the RERI expression in equation 3 as is typically done in a case-control study. With general continuous and ordinal exposures, one has to consider the magnitude of change in exposure for which one is examining the interaction. Let us consider the situation where the environmental risk factor changes from E_0 to E_1 and the genetic risk factor changes from G_0 to G_1 but other covariates z are held constant. Formally, it is defined as

$$\begin{aligned}
&\text{RERI}_{\text{OR}}(E_0, E_1, G_0, G_1) \\
&\quad = \text{OR}(G_1, E_1) - \text{OR}(G_1, E_0) - \text{OR}(G_0, E_1) + 1 \\
&\quad = \exp\{\beta_G(G_1 - G_0) + \beta_E(E_1 - E_0) \\
&\qquad + \beta_{GE}(G_1 \times E_1 - G_0 \times E_0)\} \\
&\qquad - \exp\{\beta_E(E_1 - E_0) + \beta_{GE} G_0 \times (E_1 - E_0)\} \\
&\qquad - \exp\{\beta_G(G_1 - G_0) + \beta_{GE}(G_1 - G_0) \times E_0\} + 1 \\
&\quad = f(\beta_G, \beta_E, \beta_{GE}) \approx \text{RERI}\,(E_0, E_1, G_0, G_1). \qquad (9)
\end{aligned}$$

This last approximation of risk ratios by odds ratios holds when the outcome is rare in each stratum defined by the 2 exposures or when controls are selected from the entire population, not just the noncases (27). More generally, if G and E are both categorical factors with I and J levels with coefficients corresponding to different levels of each factor, then β_G, β_E, and β_{GE} in equation 9 become $(I - 1)$-, $(J - 1)$-, and $[(I - 1)(J - 1)]$-dimensional vectors instead of scalars. Note that $\widehat{\text{RERI}}_{\text{UML}} = f(\widehat{\boldsymbol{\beta}}_{\text{UML}})$ and $\widehat{\text{RERI}}_{\text{CML}} = f(\widehat{\boldsymbol{\beta}}_{\text{CML}})$ can be viewed as functions of UML and CML estimates of relative risk parameters, where f is the function in equation 9. The variance of $\widehat{\text{RERI}}_{\text{UML}}$ and $\widehat{\text{RERI}}_{\text{CML}}$ can be calculated by means of the delta method. The EB estimator of RERI is the same as in equation 7, and its estimated variance is calculated by the delta method using the joint distribution of $(\hat{\boldsymbol{\beta}}_{\text{UML}}, \hat{\boldsymbol{\beta}}_{\text{CML}})$ as proposed by Mukherjee and Chatterjee (6) (Web Appendix 6). The Wald tests for the 3 estimators are all based on the standardized Z statistic. We have provided general codes to test for RERI on GitHub (28).

Example: analysis of $G \times E$ interactions in case-control studies of ovarian cancer

Epithelial ovarian cancer is one of the most common malignancies of the female reproductive tract. Approximately 14,080 women in the United States died from ovarian cancer in 2017 (29), comprising more deaths than those from any other cancer of the female reproductive system. There are several well-established nongenetic risk factors for ovarian cancer (30–36), and recent genome-wide association studies have identified and replicated 18 genetic variants that influence disease risk (37). To this end, the Ovarian Cancer Association Consortium has undertaken an effort to study interactions focusing on the 18 confirmed single nucleotide polymorphisms (SNPs) and 7 well-established risk factors: race, history of endometriosis, first-degree family history of ovarian cancer, OCP use, parity, tubal ligation, and age. In our illustrative analysis, we focus on OCP × SNP interaction and use genetic data from 15 Ovarian Cancer Association Consortium studies (30, 38–55) that also have data on epidemiologic risk factors (see Web Table 3).

Each SNP is coded as the number of risk alleles a subject carries, and all subsequent analysis assumes this additive genetic susceptibility model. Published odds ratios for the 18 confirmed loci in Web Table 4 are from analyses presented in the Collaborative Oncological Gene-Environment Study (38, 39, 56–61). As a parsimonious and succinct way of summarizing the effects of genetic variants across all loci for each subject, we construct a "genetic risk score" (GRS) variable as the sum of the risk allele counts across all loci and a "weighted genetic risk score" (WGRS) as the weighted sum, where the weight for each individual SNP is determined by the published log odds ratio in large meta-analysis. Polygenic risk scores have been used for risk stratification in multiple $G \times E$ papers recently (3, 62). Analysis of the marginal effect for GRS and WGRS is shown in Web Table 5. Each environmental factor is coded as a categorical variable, as described in Web Table 6. The merged $G \times E$ data set has a sample size of 11,661 subjects of European ancestry, with 4,135 cases and 7,526 controls from 13 study sites (Web Table 3).

To illustrate our inference for interactions between OCP use (1 = ever and 0 = never) and genetic risk factors, we consider both single SNP × OCP interaction and W/GRS × OCP interaction (i.e., GRS or WGRS × OCP interaction). For single-SNP analysis, we consider the top 2 hits in the 18 confirmed loci—that is, rs62274042 (SNP 1) and rs10962691 (SNP 2) as reported in Web Table 4. We used additive coding for our SNP × OCP analysis. For GRS and WGRS, we use the quartiles in controls to define a categorical variable with 4 categories. The analysis model adjusts for study site and all other environmental risk factors except race.

Simulation design

In our simulation study, we first investigate the type I error, standard power at level α, and power at empirical α (empirical type I error is used to report power in situations where type I error is not maintained) of Wald tests for $\widehat{\text{RERI}}_{\text{UML}}$, $\widehat{\text{RERI}}_{\text{CML}}$, and $\widehat{\text{RERI}}_{\text{EB}}$ under various alternative values of RERI across a spectrum of scenarios, varying the strength of the G-E association,

the main effects of G and E, the minor allele frequency of G, the prevalence of exposure E, the test size, and sample sizes. We compare the power of the Wald test for $\widehat{\text{RERI}}_{\text{CML}}$ with the previously proposed LRT for additive interaction under G-E independence (21). We also explore estimation properties like the absolute relative bias and MSE of the 3 estimators, as well as those of 2 alternative proposals, $\widehat{\text{RERI}}_{\text{EB1}}$ and $\widehat{\text{RERI}}_{\text{EB2}}$. Note that both RERI and multiplicative interaction parameters are obtained from the underlying true logistic regression model

$$\text{logit}\, P(D = 1 | G, E) = \beta_0 + \beta_E E + \beta_G G + \beta_{GE} GE,$$

where $\text{RERI} = \exp(\beta_G + \beta_E + \beta_{GE}) - \exp(\beta_G) - \exp(\beta_E) + 1$ and $\psi = \exp(\beta_{GE})$, so that the value of RERI is well-defined given ψ and vice versa, once the main effect parameters $\text{OR}_G = \exp(\beta_G)$ and $\text{OR}_E = \exp(\beta_E)$ have been specified.

We set the prevalences of G and E in controls, $p_G = (0.1, 0.2, 0.3)$ and $p_E = (0.3, 0.4, 0.5)$; the main effects $\text{OR}_G = (1.1, 1.2, 1.3)$ and $\text{OR}_E = (1.3, 1.5, 1.7)$; the sample size $N_0 = N_1 = (4{,}000, 20{,}000)$; the size of test $\alpha = (0.05, 5 \times 10^{-6})$; a change in the strength of the G-E association, $\exp(\theta_{GE})$, from 0.8 to 1.2 at a grid of 0.1; and a change in the RERI from 0 to 1.5 with a grid of 0.1. The number of simulated data sets is 1,000 when $\alpha = 0.05$, and it is 10^6 when $\alpha = 5 \times 10^{-6}$. The population parameters of cell probability $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ are defined by solving the equations in Web Appendix 7 (9, 63).

We generate data independently from the 2 multinomial distributions corresponding to the case and control populations, according to the above probabilities, with numbers of cases and controls as N_0 and N_1, respectively. We also consider another simulation setting to mimic a large-scale genome-wide search of interactions where we use random distribution for the parameters corresponding to the set of null markers. We first compute the UML, CML, and EB estimators using equations 4, 5, and 7 and then compare their type I error, power, power at empirical α, absolute relative bias, MSE, and type I error over 1,000 replications. Power is estimated by the proportion of null

Table 1. Estimates of Ovarian Cancer Risk for Interactions Between Either Single Nucleotide Polymorphisms or Genetic Risk Scores and Use of Oral Contraceptive Pills on Both the Multiplicative and Additive Scales[a]

Interaction	Multiplicative (ψ)			Additive (RERI)		
	Estimate	95% CI	*P* Value[b]	Estimate[c]	95% CI	*P* Value[b]
SNP1[d] × OCP[e]						
UML	0.94	0.73, 1.22	0.645	−0.25	−0.60, 0.10	0.162
CML	1.06	0.88, 1.28	0.548	−0.09	−0.33, 0.14	0.432
EB	1.00	0.78, 1.29	0.970	−0.16	−0.50, 0.18	0.348
SNP2[d] × OCP						
UML	0.93	0.82, 1.05	0.255	0.08	−0.18, 0.34	0.552
CML	0.94	0.85, 1.04	0.224	0.03	−0.18, 0.25	0.757
EB	0.94	0.85, 1.04	0.222	0.04	−0.11, 0.18	0.598
GRS[c] × OCP						
UML	0.82	0.65, 1.02	0.073	−0.64	−1.01, −0.27	0.001
CML	0.92	0.77, 1.08	0.305	−0.43	−0.68, −0.18	0.001
EB	0.86	0.69, 1.07	0.197	−0.54	−0.93, −0.16	0.005
WGRS[c] × OCP						
UML	0.90	0.76, 1.06	0.212	−0.61	−0.99, −0.23	0.002
CML	0.95	0.83, 1.08	0.417	−0.40	−0.67, −0.14	0.003
EB	0.93	0.81, 1.08	0.366	−0.52	−0.91, −0.13	0.009

Abbreviations: CI, confidence interval; CML, constrained maximum likelihood; EB, empirical Bayes; GRS, genetic risk score; RERI, relative excess risk due to interaction; SNP, single nucleotide polymorphism; UML, unconstrained maximum likelihood; WGRS, weighted genetic risk score.

[a] The analysis was based on subjects with European ancestry, using data on 4,135 cases and 7,526 controls from 13 study sites in the Ovarian Cancer Association Consortium (30, 38–55). The model adjusted for history of endometriosis, first-degree family history of ovarian cancer, parity, tubal ligation, age, and study site.

[b] Wald test *P* value.

[c] W/GRS is a categorical variable defined by the quartiles of WGRS in controls—for example, W/GRS = 3 if it is at or above the 75th percentile in controls and 0 if it is below the 25th percentile in controls. The minimal, 25th, 50th, and 75th percentiles and the maximum are 3, 11, 12, 14, and 22 for GRS and 0.32, 1.33, 1.53, 1.75, and 2.86 for WGRS, respectively. In this table, we present only the coefficient for the interaction term corresponding to a change in OCP from 0 to 1 and a change in WGRS from 0 to 3.

[d] SNP1 denotes rs62274042 and SNP2 denotes rs10962691. Marginal disease odds ratios corresponding to these SNPs are 1.45 (95% CI: 1.37, 1.54) and 1.25 (95% CI: 1.20, 1.30), respectively.

[e] OCP = 1 if the individual had ever used OCPs and 0 if she had never used OCPs.

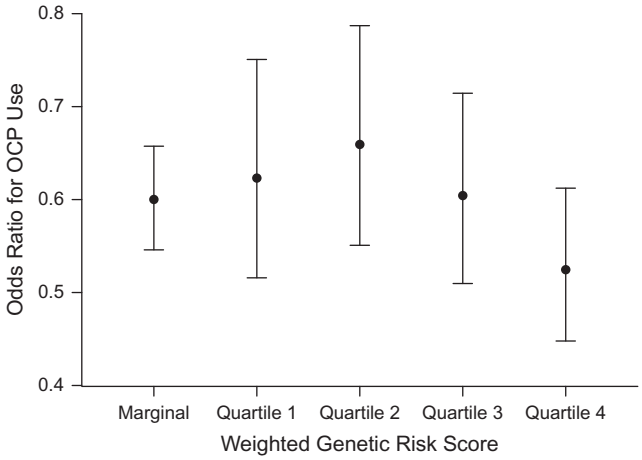

Figure 1. Odds ratio for use of oral contraceptive pills (OCPs) by quartile of weighted genetic risk score in a simulated analysis of gene-environment interaction and ovarian cancer risk. Data were obtained from 15 studies in the Ovarian Cancer Association Consortium (30, 38–55). The odds ratios were estimated using standard logistic regression with adjustment for history of endometriosis, first-degree family history of ovarian cancer, parity, tubal ligation, age, and study site. Bars, 95% confidence intervals.

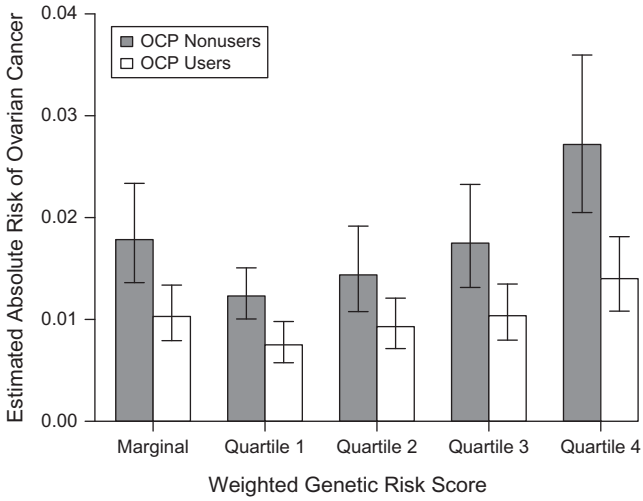

Figure 2. Predicted probability of ovarian cancer among oral contraceptive pill (OCP) users and nonusers by quartile of weighted genetic risk score in a simulated analysis of gene-environment interaction and ovarian cancer risk. Data were obtained from 15 studies in the Ovarian Cancer Association Consortium (30, 38–55). The relative risk parameters were obtained from a standard logistic regression model adjusting for history of endometriosis, first-degree family history of ovarian cancer, parity, tubal ligation, age, and study site. We assumed that approximately 1.3% of women would be diagnosed with ovarian cancer at some point during their lifetime and that 70% of women would use OCPs at some point in their life. The predicted probabilities were estimated by fixing other covariates at their most frequent value. Bars, 95% confidence intervals.

hypothesis H_0: RERI $= 0$ rejected at the given level of significance α—that is, the proportion of times $|Z| > Z_{1-\alpha/2}$, where Z is the Wald test statistic. Power at empirical α is a modified power which utilizes an empirical P value threshold as the rejection rule to control the type I error around the given significance level when the type I error at the desired nominal level is not maintained. The absolute relative bias is calculated by averaging $|\widehat{\text{RERI}} - \text{RERI}|/\text{RERI}$, and MSE is calculated by averaging $(\widehat{\text{RERI}} - \text{RERI})^2$.

RESULTS

Ovarian cancer data example

The distributions of GRS and WGRS in cases and controls are displayed in Web Figure 3. Relative to the control distributions, the upper tails of the case distributions are shifted slightly rightward. We calculate UML, CML, and EB estimators of interactions on both the multiplicative and additive scales. The estimates, corresponding confidence intervals, and Wald test P values are shown in Table 1. In SNP1 × OCP analysis, the strength of G-E association is modest: $\exp(\theta_{GE}) = 1.07$ (95% confidence interval (CI): 0.94, 1.21); the EB estimate of RERI is −0.16 (95% CI: −0.50, 0.18), where the weight on $\widehat{\text{RERI}}_{\text{UML}}$ is 43%. In SNP2 × OCP analysis, the G-E association seems weaker, with $\exp(\theta_{GE}) = 0.96$ (95% CI: 0.83, 1.11). The EB estimate of RERI is 0.04 (95% CI: −0.11, 0.18), with its weight on $\widehat{\text{RERI}}_{\text{UML}}$ decreasing to 11%. The confidence intervals corresponding to $\widehat{\text{RERI}}_{\text{EB}}$ are narrower than the corresponding intervals for $\widehat{\text{RERI}}_{\text{UML}}$. The point estimate $\widehat{\text{RERI}}_{\text{EB}}$ lies between $\widehat{\text{RERI}}_{\text{UML}}$ and $\widehat{\text{RERI}}_{\text{CML}}$, reflecting the combined efficiency-robustness feature of the EB estimator. In WGRS × OCP analysis, we evaluate interactions associated with a change in OCP use from 0 to 1 (ever users to never users) and a change in WGRS from the lowest quartile to the highest quartile (as defined through the distribution of WGRS in controls). The multiplicative measure of interaction $\hat{\psi}_{\text{EB}}$ is not significant at $\alpha = 0.05$, but $\widehat{\text{RERI}}_{\text{EB}}$ departs from 0 significantly, with an estimate of −0.52 (95% CI: −0.91, −0.13) and a very small P value ($P = 0.009$).

To visually present the results, we fit a standard logistic regression model including the main effects of OCP use and quartiles of WGRS as a categorical factor, and an interaction term for WGRS × OCP adjusting for study sites and other risk factors. Figure 1 shows the odds ratio for OCP and its corresponding 95% confidence interval according to WGRS. The odds ratio for OCP is 0.61 (95% CI: 0.50, 0.74) in the lowest WGRS quartile and 0.51 (95% CI: 0.43, 0.60) in the highest quartile. The overlapping confidence intervals indicate a nonsignificant multiplicative interaction. Additionally, if we assume that approximately 1.3% of women will be diagnosed with ovarian cancer at some point during their lifetime (29) and that 70% of women will use OCPs at some point in their life in this population (estimated from the Ovarian Cancer Association Consortium data), we can calculate the estimated lifetime risk of ovarian cancer and its corresponding 95% confidence interval within each WGRS stratum (Figure 2) for OCP users and nonusers. Estimates of lifetime absolute risk are 0.75% (95% CI: 0.57, 0.98) for OCP users and 1.23% (95% CI: 1.00, 1.51) for OCP nonusers in the lowest WGRS stratum, with a

Table 2. Empirical Familywise Type I Error Rate at a 5% Overall Level of Significance and Expected Number of False-Positive Findings Corresponding to UML, CML, and the EB Wald Test

	Proportion of Markers Satisfying Gene-Environment Independence (p_{ind})[a]					
	0.95	0.99	0.995	0.9975	0.9995	1.00
Empirical familywise type I error[b]						
UML	0.084	0.072	0.062	0.071	0.041	0.058
CML	1.000	0.994	0.966	0.745	0.874	0.064
EB	0.138	0.056	0.045	0.038	0.042	0.035
Expected no. of false positives[c]						
UML	0.085	0.073	0.062	0.071	0.042	0.059
CML	23.451	3.761	2.814	1.050	0.937	0.067
EB	0.150	0.060	0.045	0.039	0.044	0.035

Abbreviations: CML, constrained maximum likelihood; EB, empirical Bayes; RERI, relative excess risk due to interaction; SD, standard deviation; UML, unconstrained maximum likelihood.

[a] The population-level gene-environment (*G-E*) association structure among null loci is assumed to be of the form of a mixture distribution reflecting the fact that a large fraction (i.e., p_{ind}) of the single nucleotide polymorphisms are indeed independent of *E* in the population, whereas the remaining $(1 - p_{ind})$ single nucleotide polymorphisms show some departures from the independence assumption following an N (0, SD = ln(1.5)/2) distribution.

[b] The Wald test is for RERI = 0 under a large-scale genome-wide $G \times E$ scan simulation scenario with 10,000 markers and 2,000 cases and controls. Empirical familywise type I error is estimated as the empirical proportion of data sets declaring at least 1 null marker to be significant using the level of significance α/10,000. This estimates the probability of at least 1 false-positive discovery under the global null hypothesis.

[c] The expected number of false positives is estimated as the average number of falsely rejected null hypotheses, averaged over 1,000 data sets.

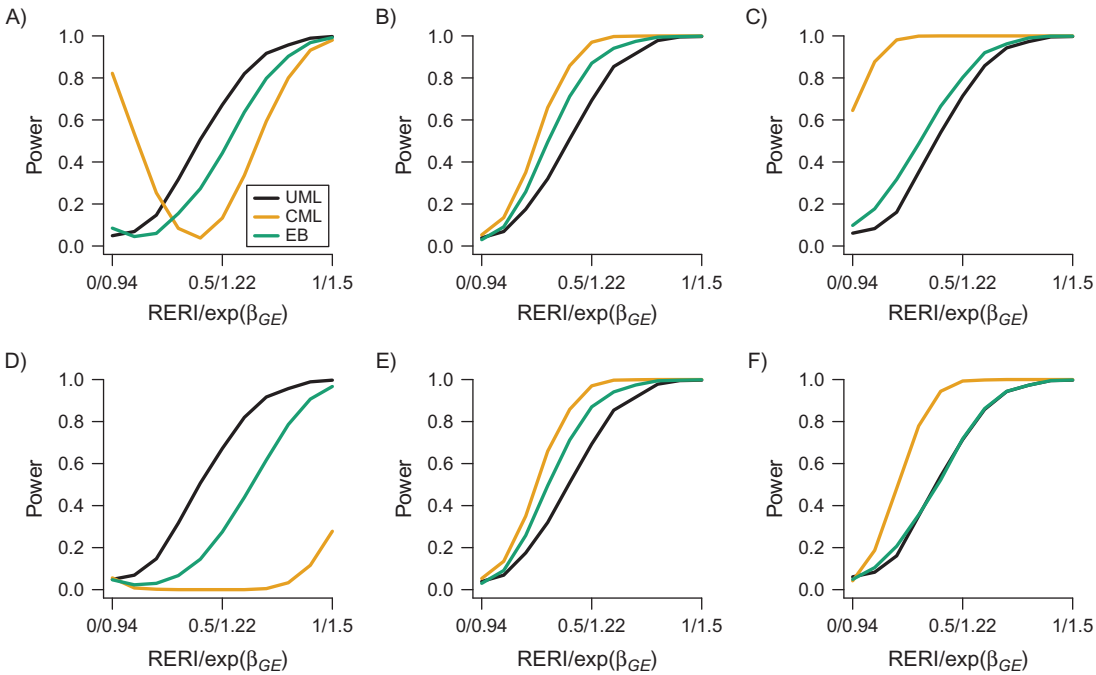

Figure 3. Power curves of unconstrained maximum likelihood (UML), constrained maximum likelihood (CML), and the empirical Bayes (EB) Wald test for the relative excess risk due to interaction (RERI) under different strengths of gene-environment (*G-E*) association in a simulated analysis of gene-environment interaction and ovarian cancer risk. Data were obtained from 15 studies in the Ovarian Cancer Association Consortium (30, 38–55). Data were generated on 4,000 cases and 4,000 controls with the fixed parameters $p_G = 0.2$, $p_E = 0.3$, $OR_G = 1.2$, and $OR_E = 1.5$. RERI changes from 0 to 1.5 with a grid level of 0.1, and corresponding multiplicative interaction changes from 0.94 to 1.78. The top panels (A, B, and C) correspond to the raw power, whereas the bottom panels (D, E, and F) correspond to the power at empirical α. The left, center, and right panels correspond to different values of the *G-E* association odds ratio (OR)—that is, $\exp(\theta_{GE}) = 0.8$, $\exp(\theta_{GE}) = 1.0$, and $\exp(\theta_{GE}) = 1.2$, respectively.

difference of 0.48% (95% CI: 0.02, 0.94); for subjects in the highest WGRS stratum, the corresponding numbers are 1.40% (95% CI: 1.08, 1.81) and 2.72% (95% CI: 2.05, 3.60), respectively, with a difference of 1.32% (95% CI: 0.24, 2.52), showing why the test for RERI is significant.

Results from the simulation study

Type I error. Web Table 7 presents type I errors for different tests of RERI. One can observe that UML maintains a nominal level α across different choices of θ_{GE}. An inflated type I error associated with CML is observed when the *G-E* independence assumption is violated. The EB test is valid when $\exp(\theta_{GE}) = 1$ and modestly inflates type I error when *G* is associated with *E*. The maximal observed type I error of EB at $\alpha = 0.05$ is 0.099 when the sample size is 40,000, the test size is 0.05, and $\exp(\theta_{GE}) = 1.1$. Web Figure 4 shows how type I error varies with $\exp(\theta_{GE})$ for the 3 estimators. The type I error of CML is very sensitive to the *G-E* association, but the performance of EB is relatively robust, with a marked reduction in type I error compared with CML. The findings remain similar for different choices of p_G, p_E, OR_G, and OR_E (Web Tables 8 and 9).

Results from additional simulation mimicking a genome-wide association study. To justify the use of the EB estimator in genome-wide assessment of *G-E* interaction, we conduct another simulation study similar to that in Mukherjee et al. (8), which generates 2,000 cases and controls with 1 causal marker together with $M - 1$ null markers, where *M* is 10,000. The *G-E* independence parameter θ_{GE} in controls has a random mixture distribution with a point mass around independence, and p_{ind} is the proportion of null loci that follow *G-E* independence. The detailed simulation setting is presented in Web Appendix 8. The expected nominal level for both the familywise error rate and the expected false-positive rate is 0.05 when *G-E* independence holds. However, if there is *G-E* dependence for a proportion of markers, Bonferroni correction cannot guarantee the nominal level for EB and CML. As shown in Table 2, when 99% of the markers are independent, EB maintains a familywise type I error rate of 0.06 and an expected number of false positives of 0.06. The performance of CML is significantly worse, with a familywise error rate of 99% and an expected number of false positives of 3.76.

Power. Figure 3 shows the power curves of the Wald test for 3 estimators with H_0: RERI = 0 under different strengths of *G-E* association (Web Tables 10–15). It is hard to compare the estimated powers directly from the figure, as the inflated type I error of CML and EB leads to misleadingly high power values. Hence, we assess the power at empirical α for CML and EB, which controls the corresponding type I error at 0.05. UML is most efficient when $\exp(\theta_{GE}) = 0.8$, CML is most efficient when $\exp(\theta_{GE}) = 1$ and 1.2, and EB power always lies in between. For a sample numerical comparison, let us compare the powers of the 3 approaches at RERI = 0.5 to represent one typical scenario. When $\exp(\theta_{GE}) = 0.8$, the empirical power of EB (0.275) is 41% lower than that of UML (0.672); meanwhile, CML has nearly 0 power. When $\exp(\theta_{GE}) = 1$, the empirical power of EB (0.870) is 25% higher than that of UML (0.693) but 10% lower than that of CML (0.970). When $\exp(\theta_{GE}) = 1.2$, the empirical power of EB (0.718) is slightly higher than that of UML (0.714) but 28% lower than that of CML (0.993). We then compare the power of the Wald test for $\widehat{\mathrm{RERI}}_{\mathrm{CML}}$ with that of the LRT for additive interaction (see Web Figure 5). The power of the LRT is uniformly slightly higher than that of the Wald test, with a true value of RERI varying from 0 to 0.5 with a grid of 0.1. Results for absolute relative bias and MSE are shown in Web Appendix 9, Web Tables 16–19, and Web Figure 6.

DISCUSSION

In this paper, we extend the EB estimator of $G \times E$ interaction proposed earlier on the multiplicative scale to the additive scale in case-control studies. The EB estimator exploits the *G-E* independence assumption to perform a tradeoff between bias and efficiency. The simulation study showed that the test based on the EB estimator can provide good control of type I error and that it is always intermediate between UML and CML with respect to power, relative bias, and mean squared error. In the ovarian cancer data example, we conducted a W/GRS × OCP analysis to illustrate the application of the proposed method. We found a significant additive W/GRS × OCP interaction but insignificant multiplicative interaction at $\alpha = 0.05$.

As an inherent limitation of case-control studies, only the relative risk can be estimated (e.g., RERI) instead of the underlying direct measure (e.g., p_{additive} in equation 1), because p_{11} can only be estimated from cohort data. However, general population incidence data from cohort studies can be combined with case-control risk-factor models to estimate absolute risks in population-based case-control studies (64), as we carried out in Figure 2. If the rare disease assumption for each configuration of *G* and *E* does not hold, approximating the relative risk by means of the odds ratio in case-control studies will not be accurate, and thus the proposed estimate of RERI may depart from the truth. By using the retrospective maximum likelihood estimates, prior guesses for disease prevalence, and adaptive combinations like the EB procedure, we can make our inference less biased under violation of the rare disease and *G-E* independence assumptions.

There is increasingly more interest in inference for additive interaction using case-control data. Tchetgen Tchetgen et al. (65) described a general approach to test for $G \times E$ additive interaction exploiting *G-E* independence which is robust to possible misspecification of main effects in the outcome regression. Han et al. (66) proposed a score test for UML and CML estimators of genetic associations under the additive null hypothesis. In the future, it will be of analytical interest to establish an EB version of the adaptive score test and the adaptive LRT, since most of the recent work has been in terms of combining estimators but not tests.

ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan (Gang Liu, Bhramar Mukherjee, Seunggeun Lee); Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California (Alice W. Lee,

Anna H. Wu, Simon A. Gayther, Celeste Leigh Pearce); Cancer Prevention and Control Research Program, Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey (Elisa V. Bandera); Department of Virus, Lifestyle and Genes, Danish Cancer Society Research Center, Copenhagen, Denmark (Allan Jensen, Susanne K. Kjaer); Program in Epidemiology, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington (Mary Anne Rossing); Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington (Mary Anne Rossing); Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, New York (Kirsten B. Moysich); Division of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany (Jenny Chang-Claude); University Cancer Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany (Jenny Chang-Claude); Department of Epidemiology, Geisel School of Medicine at Dartmouth, Dartmouth College, Hanover, New Hampshire (Jennifer A. Doherty); Gynaecological Cancer Research Centre, Women's Cancer, Institute for Women's Health, University College London, London, United Kingdom (Aleksandra Gentry-Maharaj, Usha Menon); Radboud University Medical Center, Radboud Institute for Health Sciences, Nijmegen, the Netherlands (Lambertus Kiemeney); Department of Obstetrics, Gynecology, and Reproductive Sciences, Division of Gynecologic Oncology, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania (Francesmary Modugno); Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania (Francesmary Modugno); Womens Cancer Research Program, Magee-Womens Research Institute and University of Pittsburgh Cancer Institute, Pittsburgh, Pennsylvania (Francesmary Modugno); Department of Obstetrics and Gynaecology, Radboud University Medical Center, Nijmegen, the Netherlands (Leon Massuger); Department of Health Sciences Research, Division of Epidemiology, Mayo Clinic, Rochester, Minnesota (Ellen L. Goode); University of Kansas Medical Center, Kansas City, Kansas (Brooke L. Fridley); Obstetrics and Gynecology Center, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts (Kathryn L. Terry, Daniel W. Cramer); Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts (Kathryn L. Terry, Daniel W. Cramer); School of Women's and Children's Health, University of New South Wales, Sydney, New South Wales, Australia (Susan J. Ramus); Kinghorn Cancer Centre, Garvan Institute of Medical Research, Sydney, New South Wales, Australia (Susan J. Ramus); Genetic Epidemiology Research Institute, Center for Cancer Genetics Research and Prevention, School of Medicine, University of California, Irvine, Irvine, California (Hoda Anton-Culver, Argyrios Ziogas); Strangeways Research Laboratory, Department of Public Health and Primary Care, School of Clinical Medicine, University of Cambridge, Cambridge, United Kingdom (Jonathan P. Tyrer, Paul D. Pharoah); Department of Public Health Sciences, School of Medicine, University of Virginia, Charlottesville, Virginia (Joellen M. Schildkraut); Department of Gynecology, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark (Susanne K. Kjaer); QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia (Penelope M. Webb); Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, University of Texas, Houston, Texas (Roberta B. Ness); Department of Obstetrics and Gynecology, Duke University Medical Center, Durham, North Carolina (Andrew Berchuck); Department of Oncology, Center for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, United Kingdom (Paul D. Pharoah); Department of Chronic Disease Epidemiology, School of Public Health, Yale University, New Haven, Connecticut (Harvey Risch); and Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan (Celeste Leigh Pearce).

This work was supported by the National Cancer Institute, US National Institutes of Health (grant R01 CA076016), and the National Cancer Institute's Genetic Associations and Mechanisms in Oncology (GAME-ON) initiative (grant U19-CA148112). It was also supported by the National Institutes of Health (grants P30 CA14089, R01 CA61132, P01 CA17054, N01 PC67010, R03 CA113148, N01 CN025403, and R03 CA115195 (USC), K07 CA095666, R01 CA83918, K22 CA138563, and P30 CA072720 (NJO), R01 CA122443, P30 CA15083, and P50 CA136393R01 (MAY), R01 CA112523 and R01 CA87538 (DOV), R01 CA058860 (UCI), R01 CA063678, R01 CA074850, and R01 CA080742 (CON), R01 CA76016 (NCO), R01 CA54419 and P50 CA105009 (NEC), R01 CA61107 (MAL), and R01 CA095023, R01 CA126841, M01 RR000056, P50 CA159981, and K07 CA80668 (HOP)); the California Cancer Research Program (grants 0001389V20170 and 2110200 (USC)); the German Federal Ministry of Education and Research, Program of Clinical Biomedical Research (grant 01GB9401 (GER)); the German Cancer Research Centre (GER); the Danish Cancer Society (grant 94 222 52 (MAL)); Mermaid I (MAL); the Eve Appeal/Oak Foundation (UKO); the Cancer Institute of New Jersey (NJO); the National Institute for Health Research University College London Hospitals Biomedical Research Centre (UKO); the US Army Medical Research and Materiel Command (grants W81XWH-10-1-02802 (NEC), DAMD17-02-1-0669 (HOP), DAMD17-02-1-0666 (NCO), and DAMD17-01-1-0729 (AUS)); the Roswell Park Alliance Foundation (HOP); the Cancer Councils of New South Wales, Victoria, Queensland, South Australia, and Tasmania (Multi-State Application numbers 191, 211, and 182 (AUS)); the Cancer Foundation of Western Australia (AUS); the National Health and Medical Research Council of Australia (grants 199600 and 400281 (AUS)); the Mayo Foundation (MAY); the Minnesota Ovarian Cancer Alliance (MAY); the Fred C. and Katherine B. Andersen Foundation (MAY); Radboud University Medical Centre (NTH); the Lon V Smith Foundation (grant LVS-39420 (UCI)); the National Institute of Environmental Health Sciences, US National Institutes of Health (grant T32 ES013678 to A.W.L.); and the National Health and Medical Research Council of Australia (fellowship 1043134 to P.M.W.). (See Web Table 3 for definitions of parenthetical study abbreviations.) The research was also supported by the National Cancer Institute (grant P30 CA046592). Lastly, this work was also supported by the National Science Foundation (grant NSF DMS 1406712) and the National Institute of Environmental Health Sciences (grant NIH ES 20811).

The Collaborative Oncological Gene-Environment Study is funded through the European Commission's Seventh Framework Programme (agreement 223175 HEALTH F2 2009-223175). The Ovarian Cancer Association Consortium is supported by a grant from the Ovarian Cancer Research Fund thanks to donations by the family and friends of Kathryn Sladek Smith (grant PPD/RPCI.07).

We thank all of the researchers, clinicians, and technical and administrative staff who have made possible the many studies contributing to this work. In particular, we thank Dr. D. Bowtell, Dr. A. DeFazio, Dr. D. Gertig, Dr. A. Green, Dr. P. Parsons, Dr. N. Hayward, and Dr. D. Whiteman (AUS); the staff of the genotyping unit, Dr. S. LaBoissiere and F. Robidoux (Génome Québec, Montreal, Quebec, Canada); Dr. U. Eilber (GER); and Dr. I. Jacobs, Dr. M. Widschwendter, Dr. E. Wozniak, N. Balogun, Dr. A. Ryan, and J. Ford (UKO).

The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest: none declared.

REFERENCES

1. Hutter CM, Chang-Claude J, Slattery ML, et al. Characterization of gene-environment interactions for colorectal cancer susceptibility loci. *Cancer Res*. 2012;72(8):2036–2044.
2. Hsu L, Jiao S, Dai JY, et al. Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genet Epidemiol*. 2012;36(3):183–194.
3. Garcia-Closas M, Rothman N, Figueroa J, et al. Common genetic polymorphisms modify the effect of smoking on absolute risk of bladder cancer. *Cancer Res*. 2013;73(7):2211–2220.
4. Figueiredo JC, Hsu L, Hutter CM, et al. Genome-wide diet-gene interaction analyses for risk of colorectal cancer. *PLoS Genet*. 2014;10(4):e1004228.
5. Lewinger JP, Morrison JL, Thomas DC, et al. Efficient two-step testing of gene-gene interactions in genome-wide association studies. *Genet Epidemiol*. 2013;37(5):440–451.
6. 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.
7. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol*. 2009;169(2):219–226.
8. Mukherjee B, Ahn J, Gruber SB, et al. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *Am J Epidemiol*. 2012;175(3):177–190.
9. Boonstra PS, Mukherjee B, Gruber SB, et al. Tests for gene-environment interactions and joint effects with exposure misclassification. *Am J Epidemiol*. 2016;183(3):237–247.
10. Thomas D. Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annu Rev Public Health*. 2010;31:21–36.
11. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979;66(3):403.
12. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med*. 1994;13(2):153–162.
13. Umbach DM, Weinberg CR. Designing and analyzing case-control studies to exploit independence of genotype and exposure. *Stat Med*. 1997;16(15):1731–1743.
14. Du M, Zhang X, Hoffmeister M, et al. No evidence of gene-calcium interactions from genome-wide analysis of colorectal cancer risk. *Cancer Epidemiol Biomarkers Prev*. 2014;23(12):2971–2976.
15. Joshi A, 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.
16. Greenland S, Lash TL, Rothman KJ. Concepts of interaction. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Wolters Kluwer Health; 2008:71–87.
17. VanderWeele TJ. A word and that to which it once referred: assessing "biologic" interaction. *Epidemiology*. 2011;22(4):612–613.
18. Hosmer DW, Lemeshow S. Confidence interval estimation of interaction. *Epidemiology*. 1992;3(5):452–456.
19. Zou G. On the estimation of additive interaction by use of the four-by-two table and beyond. *Am J Epidemiol*. 2008;168(2):212–224.
20. VanderWeele TJ. Sample size and power calculations for additive interactions. *Epidemiol Methods*. 2012;1(1):159–188.
21. 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.
22. Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*. 2005;92(2):399–418.
23. VanderWeele TJ. An introduction to interaction analysis. In: *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York, NY: Oxford University Press; 2015:255–260.
24. Zhang L, Mukherjee B, Ghosh M, et al. Accounting for error due to misclassification of exposures in case-control studies of gene-environment interaction. *Stat Med*. 2008;27(15):2756–2783.
25. Chen YH, Chatterjee N, Carroll R. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *J Am Stat Assoc*. 2009;104(485):220–233.
26. Leeb H, Pötscher BM. Sparse estimators and the oracle property, or the return of Hodges' estimator. *J Econom*. 2008;142(1):201–211.
27. Knol MJ, Vandenbroucke JP, Scott P, et al. What do case-control studies estimate? Survey of methods and assumptions in published case-control research. *Am J Epidemiol*. 2008;168(9):1073–1081.
28. Greyson L. UML, CML and EB Wald tests of G-E interaction in multiplicative and additive scale. https://github.com/GreysonL/RERI/releases. Published January 12, 2017. Accessed May 12, 2017.
29. Surveillance, Epidemiology, and End Results Program, National Cancer Institute. Cancer Stat Facts: ovarian cancer. http://seer.cancer.gov/statfacts/html/ovary.html. Published April 14, 2017. Accessed October 27, 2017.
30. Collaborative Group on Epidemiological Studies of Ovarian Cancer, Beral V, Doll R, et al. Ovarian cancer and oral contraceptives: collaborative reanalysis of data from 45 epidemiological studies including 23,257 women with ovarian cancer and 87,303 controls. *Lancet*. 2008;371(9609):303–314.

31. Pike MC, Pearce CL, Peters R, et al. Hormonal factors and the risk of invasive ovarian cancer: a population-based case-control study. *Fertil Steril*. 2004;82(1):186–195.
32. Whiteman DC, Murphy MF, Cook LS, et al. Multiple births and risk of epithelial ovarian cancer. *J Natl Cancer Inst*. 2000; 92(14):1172–1177.
33. Tung KH, Goodman MT, Wu AH, et al. Reproductive factors and epithelial ovarian cancer risk by histologic type: a multiethnic case-control study. *Am J Epidemiol*. 2003;158(7):629–638.
34. Cibula D, Widschwendter M, Májek O, et al. Tubal ligation and the risk of ovarian cancer: review and meta-analysis. *Hum Reprod Update*. 2011;17(1):55–67.
35. Pearce CL, Templeman C, Rossing MA, et al. Association between endometriosis and risk of histological subtypes of ovarian cancer: a pooled analysis of case-control studies. *Lancet Oncol*. 2012;13(4):385–394.
36. Auranen A, Pukkala E, Mäkinen J, et al. Cancer incidence in the first-degree relatives of ovarian cancer patients. *Br J Cancer*. 1996;74(2):280–284.
37. Pearce CL, Rossing M, Lee AW, et al. Combined and interactive effects of environmental and GWAS-identified risk factors in ovarian cancer. *Cancer Epidemiol Biomarkers Prev*. 2013;22(5):880–890.
38. Bolton KL, Tyrer J, Song H, et al. Common variants at 19p13 are associated with susceptibility to ovarian cancer. *Nat Genet*. 2010;42(10):880–884.
39. Goode EL, Chenevix-Trench G, Song H, et al. A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nat Genet*. 2010;42(10):874–879.
40. Merritt MA, Green AC, Nagle CM, et al. Talcum powder, chronic pelvic inflammation and NSAIDs in relation to risk of epithelial ovarian cancer. *Int J Cancer*. 2008;122(1): 170–176.
41. Risch HA, Bale AE, Beck PA, et al. PGR +331 A/G and increased risk of epithelial ovarian cancer. *Cancer Epidemiol Biomarkers Prev*. 2006;15(9):1738–1741.
42. Rossing MA, Cushing-Haugen KL, Wicklund KG, et al. Menopausal hormone therapy and risk of epithelial ovarian cancer. *Cancer Epidemiol Biomarkers Prev*. 2007;16(12):2548–2556.
43. Royar J, Becher H, Chang-Claude J. Low-dose oral contraceptives: protective effect on ovarian cancer risk. *Int J Cancer*. 2001;95(6):370–374.
44. Goodman MT, Lurie G, Thompson PJ, et al. Association of two common single-nucleotide polymorphisms in the *CYP19A1* locus and ovarian cancer risk. *Endocr Relat Cancer*. 2008;15(4):1055–1060.
45. Ness RB, Dodge RC, Edwards RP, et al. Contraception methods, beyond oral contraceptives and tubal ligation, and risk of ovarian cancer. *Ann Epidemiol*. 2011;21(3):188–196.
46. Glud E, Kjaer SK, Thomsen BL, et al. Hormone therapy and the impact of estrogen intake on the risk of ovarian cancer. *Arch Intern Med*. 2004;164(20):2253–2259.
47. Goode EL, Maurer MJ, Sellers TA, et al. Inherited determinants of ovarian cancer survival. *Clin Cancer Res*. 2010;16(3):995–1007.
48. Kelemen LE, Sellers TA, Schildkraut JM, et al. Genetic variation in the one-carbon transfer pathway and ovarian cancer risk. *Cancer Res*. 2008;68(7):2498–2506.
49. Schildkraut JM, Iversen ES, Wilson MA, et al. Association between DNA damage response and repair genes and risk of invasive serous ovarian cancer. *PLoS One*. 2010;5(4):e10061.
50. Schildkraut JM, Moorman PG, Bland AE, et al. Cyclin E overexpression in epithelial ovarian cancer characterizes an etiologic subgroup. *Cancer Epidemiol Biomarkers Prev*. 2008; 17(3):585–593.
51. Terry KL, De Vivo I, Titus-Ernstoff L, et al. Androgen receptor cytosine, adenine, guanine repeats, and haplotypes in relation to ovarian cancer risk. *Cancer Res*. 2005;65(13):5974–5981.
52. Bandera EV, King M, Chandran U, et al. Phytoestrogen consumption from foods and supplements and epithelial ovarian cancer risk: a population-based case control study. *BMC Womens Health*. 2011;11:40.
53. Ziogas A, Gildea M, Cohen P, et al. Cancer risk estimates for family members of a population-based family registry for breast and ovarian cancer. *Cancer Epidemiol Biomarkers Prev*. 2000;9(1):103–111.
54. Balogun N, Gentry-Maharaj A, Wozniak EL, et al. Recruitment of newly diagnosed ovarian cancer patients proved challenging in a multicentre biobanking study. *J Clin Epidemiol*. 2011;64(5):525–530.
55. Wu AH, Pearce CL, Tseng CC, et al. Markers of inflammation and risk of ovarian cancer in Los Angeles County. *Int J Cancer*. 2009;124(6):1409–1415.
56. Kuchenbaecker KB, Ramus SJ, Tyrer J, et al. Identification of six new susceptibility loci for invasive epithelial ovarian cancer. *Nat Genet*. 2015;47(2):164–171.
57. Song H, Ramus SJ, Tyrer J, et al. A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2. *Nat Genet*. 2009;41(9):996–1000.
58. Pharoah PD, Tsai YY, Ramus SJ, et al. GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nat Genet*. 2013;45(4):362–370, 370e1–370e2.
59. Permuth-Wey J, Lawrenson K, Shen HC, et al. Identification and molecular characterization of a new ovarian cancer susceptibility locus at 17q21.31. *Nat Commun*. 2013;4:1627.
60. Bojesen SE, Pooley KA, Johnatty SE, et al. Multiple independent variants at the *TERT* locus are associated with telomere length and risks of breast and ovarian cancer. *Nat Genet*. 2013;45(4):371–384, 384e1–384e2.
61. Couch FJ, Wang X, McGuffog L, et al. Genome-wide association study in *BRCA1* mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLoS Genet*. 2013;9(3):e1003212.
62. Li S, Zhao JH, Luan J, et al. Cumulative effects and predictive value of common obesity-susceptibility variants identified by genome-wide association studies. *Am J Clin Nutr*. 2010;91(1):184–190.
63. Mukherjee B, Ahn J, Gruber SB, et al. Tests for gene-environment interaction from case-control data: a novel study of type I error, power and designs. *Genet Epidemiol*. 2008; 32(7):615–626.
64. Risch HA, Yu H, Lu L, et al. Detectable symptomatology preceding the diagnosis of pancreatic cancer and absolute risk of pancreatic cancer diagnosis. *Am J Epidemiol*. 2015;182(1):26–34.
65. Tchetgen Tchetgen E, Sofer T, Wong BH. A general approach to detect gene (G)-environment (E) additive interaction leveraging G-E independence in case-control studies. (Harvard University Biostatistics Working Paper no. 177). http://biostats.bepress.com/harvardbiostat/paper177/. Accessed June 30, 2014.
66. 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.