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

Gang Liu, Seunggeun Lee, Alice W. Lee, Anna H. Wu, Elisa V. Bandera, Allan Jensen, Mary Anne Rossing, Kirsten B. Moysich, Jenny Chang-Claude, Jennifer Doherty, Aleksandra Gentry-Maharaj, Lambertus Kiemeney, Simon A. Gayther, Francesmary Modugno, Leon Massuger, Ellen L. Goode, Brooke 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, Celeste Leigh Pearce and Bhramar Mukherjee

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)

Author affiliations: Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA (Gang Liu, Seunggeun Lee, and Bhramar Mukherjee); Department of Preventive Medicine, University of Southern California Keck School of Medicine, Los Angeles, California, USA (Alice W. Lee, Anna H. Wu, Malcolm C. Pike, Celeste Leigh Pearce); Department of Cancer Epidemiology, Division of Population Sciences, Moffitt Cancer Center, Tampa, Florida, USA (Catherine M. Phelan); Cancer Prevention and Control, Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey, USA (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, USA (Mary Rossing); Department of Epidemiology, University of Washington, Seattle, Washington, USA (Mary Rossing); Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, New York, USA (Kirsten B. Moysich); Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany (Jennifer Chang-Claude); University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany (Jennifer Chang-Claude); Department of Epidemiology, The Geisel School of Medicine at Dartmouth, Hanover, NH, USA (Jennifer Doherty); Women's Cancer, Institute for Women's Health, University College London, London, United Kingdom (Aleksandra Gentry-Maharaj); 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, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA (Francesmary Modugno); Department of Epidemiology, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania, USA (Francesmary Modugno); Womens Cancer Research Program, Magee-Womens Research Institute and University of Pittsburgh Cancer Institute, Pittsburgh, Pennsylvania, USA (Francesmary Modugno); Radboud University Medical Center, Radboud Institute for Molecular Life Sciences, Nijmegen, The Netherlands (Leon Massuger); Department of Health Sciences Research, Division of Epidemiology, Mayo Clinic, Rochester, Minnesota, USA (Ellen L. Goode); Kansas IDeA Network of Biomedical Research Excellence Bioinformatics Core, University of Kansas Cancer Center, Kansas City, Kansas, USA (Brooke Fridley); Obstetrics and Gynecology Center, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA (Kathryn L. Terry, Daniel W. Cramer); Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA (Kathryn L. Terry, Daniel W. Cramer); Genetic Epidemiology Research Institute, UCI Center for Cancer Genetics Research and Prevention, School of Medicine, Department of Epidemiology, University of California Irvine, Irvine, California, USA (Hoda Anton-Culver, Argyrios Ziogas); Department of Public Health and Primary Care, Center for Cancer Genetic Epidemiology, University of Cambridge, Strangeways Research Laboratory, Cambridge, UK (Jonathan Tyrer, Paul D. Pharoah); Department of Public Health Sciences, The University of Virginia, Charlottesville, Virginia, USA (Joellen M. Schildkraut); Department of Gynecology, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark (Susanne K. Kjaer); Queensland Institute of Medical Research, Brisbane, Australia (Penelope M. Webb); University of Texas School of Public Health, Houston, Texas, USA (Roberta B. Ness); Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York, USA (Malcolm C. Pike); Department of Women's Cancer, EGA Institute for Women's Health, University College London, London, United Kingdom (Usha Menon); Department of

Obstetrics and Gynecology, Duke University Medical Center, Durham, North Carolina, USA (Andrew Berchuck); Department of Oncology, Center for Cancer Genetic Epidemiology, University of Cambridge, University of Cambridge, Cambridge, United Kingdom (Paul D. Pharoah); Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, Connecticut, USA (Harvey Risch); and Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, Michigan, USA (Celeste Leigh Pearce).

This work was supported by: the US National Cancer Institute (R01 CA076016); the COGS project is funded through a European Commission's Seventh Framework Programme grant (agreement number 223175 HEALTH F2 2009-223175); the Genetic Associations and Mechanisms in Oncology (GAME-ON): a NCI Cancer Post-GWAS Initiative (U19-CA148112); 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 (PPD/RPCI.07).

It was also supported by the National Institutes of Health (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]); California Cancer Research Program (0001389V20170 and 2110200 [USC]); German Federal Ministry of Education and Research of Germany, Programme of Clinical Biomedical Research (01GB9401 [GER]); German Cancer Research Centre (GER); Danish Cancer Society (94 222 52 [MAL]); Mermaid I [MAL]; Eve Appeal/Oak Foundation (UKO); Cancer Institute of New Jersey (NJO); the National Institute for Health Research University College London Hospitals Biomedical Research Centre (UKO); US Army Medical Research and Materiel Command (W81XWH-10-1-02802 [NEC], DAMD17-02-1-0669 [HOP], DAMD17-02-1-0666 [NCO], and DAMD17-01-1-0729 [AUS]); Roswell Park Alliance Foundation (HOP); Cancer Councils of New South Wales, Victoria, Queensland, South Australia and Tasmania (Multi-State Application Numbers 191, 211 and 182 [AUS]); Cancer Foundation of Western Australia (AUS); National Health and Medical Research Council of Australia (199600 and 400281 [AUS]); Mayo Foundation (MAY); Minnesota Ovarian Cancer Alliance (MAY); Fred C. and Katherine B. Andersen Foundation (MAY); Radboud University Medical Centre (NTH); Lon V Smith Foundation (LVS-39420 [UCI]); National Institute of Environmental Health Sciences (T32 ES013678 for A.W.L.); National Health and Medical Research Council of Australia (for G.C.T. and P.W.).

Research reported in this publication was partially supported by NCI award number P30 CA008748 (PI: Thompson) to Memorial Sloan Kettering Cancer Center. It was also supported by the National Cancer Institute of the National Institutes of Health (P30 CA046592). Lastly, this work was also supported by the National Science Foundation (NSF DMS 1406712) and the National Institutes of Health (NIH ES 20811). The content 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.

Running head: Adaptive Test for Additive Interaction

Abbreviations:

CML: constrained maximum likelihood;

EB: empirical Bayes;

GRS: genetic risk score;

GWAS: genome-wide association studies;

LRT: likelihood ratio test;

MLE: maximum likelihood estimates;

OCP: oral contraceptive pill;

RERI: relative excess risk due to interaction;

SNP: single nucleotide polymorphism;

UML: unconstrained maximum likelihood;

WGRS: weighted genetic risk score.

ABSTRACT

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 the 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. This paper extends the empirical Bayes (EB) approach previously developed for multiplicative interaction that trades off between bias and efficiency in a data-adaptive way, to the additive scale. An EB estimator of *Relative Excess Risk due to Interaction* is derived and the corresponding Wald test is proposed with 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 power gain compared to the standard logistic regression analysis and better control of Type I error when compared to the analysis assuming gene-environment independence. We illustrate the methods with data from the Ovarian Cancer Association Consortium.

Keywords: Bias-Variance Trade-off, Effect Modification, Empirical Bayes, Genetic Risk Score, Relative Excess Risk, Shrinkage.

Word Count: Abstract: 186 words, Body: 4460 words

INTRODUCTION

There has been an increasing interest in searching for gene by environment interaction (G x E) in the post genome-wide association studies (GWAS) era with limited success (1-5). A number of methods have been proposed for efficient search of G x 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 the environmental exposure, say, E, can potentially be modified via an intervention, the additive gene x environment 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 limited access to an intervention are available. Moreover, the additive measure of interaction corresponds more closely to the notion of mechanistic or causal measures of interaction (16, 17).

Although 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 RERI can be viewed as a function of both main effects and multiplicative interaction parameters in a logistic regression model. Standard Delta theorem can be applied to provide asymptotic variance and subsequently a Wald test for the null hypotheses RERI=0 can be conducted (18-20). The fact that RERI=0 if and only if the additive null holds provides us 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 the 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 ML method. However, it is not clear how to extend the LRT in an EB-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 (MLE) and Wald test of the RERI parameter without (unconstrained MLE) and with assuming gene-environment independence (constrained MLE). We then extend the empirical Bayes-type shrinkage approach for multiplicative G x E interaction proposed by Mukherjee et.al (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 (22). We conduct a simulation study to compare the performance of various tests and illustrate our method by applying it to study the interaction between oral contraceptive pill (OCP) use 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 (non-carrier) of the susceptible genetic marker. Let D denote the disease status, where D=1 (D=0) stands for an affected (unaffected) individual. Let N_0 and N_1 be the number 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.

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 vector 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 two independent multinomial distributions in controls and cases respectively, namely, $\boldsymbol{r_0}$~Multinomial $(N_0, \boldsymbol{p_0})$ and $\boldsymbol{r_1}$~Multinomial$(N_1, \boldsymbol{p_1})$.

Let us introduce the following notation for the key parameters of interest. Let $OR_E = \frac{P(D=1|E=1,G=0)}{P(D=0|E=1,G=0)} \Big/ \frac{P(D=1|E=0,G=0)}{P(D=0|E=0,G=0)} = p_{01}p_{12}/p_{02}p_{11}$ denote the odds ratio associated with E for non-susceptible individuals (G=0), $OR_G = \frac{P(D=1|G=1,E=0)}{P(D=0|G=1,E=0)} \Big/ \frac{P(D=1|G=0,E=0)}{P(D=0|G=0,E=0)} = p_{01}p_{13}/p_{03}p_{11}$

denote the odds ratio associated with G for unexposed individuals (E=0) and $OR_{GE} = \frac{P(D=1|E=1,G=1)}{P(D=0|E=1,G=1)}\Big/\frac{P(D=1|E=0,G=0)}{P(D=0|E=0,G=0)} = p_{01}p_{14}/p_{04}p_{11}$ denote the joint odds ratio associated with the sub-group G=1 and E=1 compared to the reference group of G=0 and E=0. The multiplicative interaction parameter ψ is defined as:

$$\psi = \frac{OR_{GE}}{OR_G OR_E} = \frac{p_{02}p_{03}p_{11}p_{14}}{p_{01}p_{04}p_{12}p_{13}} = \frac{\frac{p_{11}p_{14}}{p_{12}p_{13}}}{\exp(\theta_{GE})}, \text{where } \theta_{GE} = \log\frac{p_{01}p_{04}}{p_{02}p_{03}}.$$

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

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

$$=$$

$$\mathrm{P(D=1|E=1,G=1) - P(D=1|E=1,G=0) - P(D=1|E=0,G=1) +}$$
$$\mathrm{P(D=1|E=0,G=0)} \qquad (1)$$

Dividing (1) throughout by $\mathrm{P(D=1|E=0,G=0)}$ we obtain a new measure relative excess risk due to interaction (RERI)

$$RERI_{RR} = RR_{GE} - RR_G - RR_E + 1. \qquad (2)$$

When the disease is rare, OR approximates RR. Hence, we have

$$RERI_{OR} \approx OR_{GE} - OR_G - OR_E + 1. \qquad (3)$$

Note that by (1) and (3), testing $\mathrm{H_0}: \mathrm{p_{additive}} = 0$ is equivalent to testing $\mathrm{H_0}: RERI_{RR} = 0$, which is typically translated into $\mathrm{H_0}: RERI_{OR} = 0$ in a case-control study as described in

VanderWeele (23). After defining the above relevant parameters of interest, we use the definition of RERI in equation (3) in terms of ORs to proceed with inference under case-control sampling assuming the disease is rare for all configurations of G and E.

Unconstrained maximum likelihood estimation

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

$$\hat{\psi}_{uml} = \frac{\widehat{OR}_{GE}}{\widehat{OR}_G\widehat{OR}_E} = \frac{r_{02}r_{03}r_{11}r_{14}}{r_{01}r_{04}r_{12}r_{13}}, \qquad \hat{\sigma}^2_{uml} = Var(\log(\hat{\psi}_{uml})) = \sum_{d=0}^{1}\sum_{j=1}^{4}\frac{1}{r_{dj}}$$

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

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

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

Note that $\boldsymbol{r_0}$ and $\boldsymbol{r_1}$ are realizations from two independent multinomial distributions, and we can employ Delta method (Web Appendix 2) to obtain the asymptotic variance of $\widehat{RERI}_{\text{uml}}$, which is the same as noted in (17-19). The Wald test for interaction is based on the standardized Z statistic $Z_{uml} = \widehat{RERI}_{uml}/\sqrt{\widehat{Var}\left(\widehat{RERI}_{uml}\right)}$ which follows a 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 rare disease assumptions, Zhang et.al (24) proposed the constrained MLEs (CML) 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 $\hat{p}_{1j} = \frac{r_{1j}}{N_1}, j = 1,2,3,4$.

We obtain the corresponding OR estimates by substituting p_{dj} with its constrained MLE under G-E independence, $\widehat{OR}_E = \frac{r_{12}(r_{01}+r_{03})}{r_{11}(r_{02}+r_{04})}$, $\widehat{OR}_G = \frac{r_{13}(r_{01}+r_{02})}{r_{11}(r_{03}+r_{04})}$, $\widehat{OR}_{GE} = \frac{r_{14}(r_{01}+r_{02})(r_{01}+r_{03})}{r_{11}(r_{02}+r_{04})(r_{03}+r_{04})}$ and $\hat{\psi}_{\mathrm{cml}} = \frac{r_{11}r_{14}}{r_{12}r_{13}}$, $\hat{\sigma}_{cml}^2 = Var(\log(\hat{\psi}_{cml})) = \sum_{j=1}^{4} \frac{1}{r_{1j}}$. 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 ORs under the constraint into equation (3). Formally, the CML estimator for RERI is given by

$$\widehat{RERI}_{cml} = \frac{(r_{01}+r_{03})(r_{01}+r_{02})r_{14}}{(r_{02}+r_{04})(r_{03}+r_{04})r_{11}} - \frac{(r_{01}+r_{02})r_{13}}{(r_{03}+r_{04})r_{11}} - \frac{(r_{01}+r_{03})r_{12}}{(r_{02}+r_{04})r_{11}} + 1. \quad (5)$$

Under 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 Delta method, which is shown in Web Appendix 3. The Wald test for RERI in a constrained model again uses the standardized Z statistic $Z_{cml} = \widehat{RERI}_{cml} / \sqrt{\widehat{Var}\left(\widehat{RERI}_{cml}\right)}$, and the power of the test is slightly lower than LRT for additive interaction in (21) as will be illustrated through our simulations. Under violation of gene-environment 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 et.al (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 G-E independence assumption and makes a trade-off between bias and efficiency. Formally, the EB estimator of multiplicative interaction is given by

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

where $\hat{\psi}_{cml} = \frac{r_{11}r_{14}}{r_{12}r_{13}}$, $\hat{\psi}_{uml} = \frac{r_{02}r_{03}r_{11}r_{14}}{r_{01}r_{04}r_{12}r_{13}}$, $\hat{\sigma}^2_{uml} = \sum_{d=0}^{1}\sum_{j=1}^{4}\frac{1}{r_{dj}}$ and $\hat{\theta}_{GE} = \log\frac{r_{01}r_{04}}{r_{02}r_{03}}$.

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

$$\widehat{RERI}_{EB} = \frac{(\widehat{RERI}_{uml}-\widehat{RERI}_{cml})^2}{\widehat{Var}(\widehat{RERI}_{uml})+(\widehat{RERI}_{uml}-\widehat{RERI}_{cml})^2}\widehat{RERI}_{uml} + \frac{\widehat{Var}(\widehat{RERI}_{uml})}{\widehat{Var}(\widehat{RERI}_{uml})+(\widehat{RERI}_{uml}-\widehat{RERI}_{cml})^2}\widehat{RERI}_{cml}$$

$$= \widehat{RERI}_{uml} + K(\widehat{RERI}_{cml} - \widehat{RERI}_{uml}) \qquad (7),$$

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 $\hat{\kappa} = \widehat{RERI}_{uml} - \widehat{RERI}_{cml}$ and $V = \widehat{Var}\left(\widehat{RERI}_{uml}\right)$. To explain the intuitive rationale behind the estimator, observe that as $\widehat{\theta_{GE}} \to 0$, i.e. as the data provide the evidence in favor of G-E independence, $\widehat{RERI}_{uml} - \widehat{RERI}_{cml} \to 0$, the estimator puts more weight on CML estimator to gain more efficiency, and as $\widehat{\theta_{GE}} \to \infty$. i.e. as the G-E dependence becomes stronger in control population, $\widehat{RERI}_{uml} - \widehat{RERI}_{cml}$ becomes larger, then the EB estimator puts more weight on 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 $\widehat{RERI}_{EB}$ is derived by Delta method (See Web Appendix 4), assuming $\widehat{Var}\left(\widehat{RERI}_{uml}\right)$ as a constant relative to the order of magnitude of the point estimates (6). We use Wald test for the EB estimator based on the standardized Z statistic $Z_{EB} = \widehat{RERI}_{EB} / \sqrt{\widehat{Var}\left(\widehat{RERI}_{EB}\right)}$.

Remark 1. We also considered two other forms of adaptive weights. One is to modify the shrinkage factor K in (7) and let $\widehat{k^*} = \hat{\theta}_{GE}$ instead of $\widehat{RERI}_{uml} - \widehat{RERI}_{cml}$, namely, $\widehat{RERI}_{EB1} = \widehat{RERI}_{uml} + K^*(\widehat{RERI}_{cml} - \widehat{RERI}_{uml})$, where $K^* = V\left(V + \widehat{\kappa^*}\widehat{\kappa^*}^{T}\right)^{-1}$. The other is to plug in the EB estimates, $\widehat{OR}_{EB}$, obtained from using the retrospective likelihood framework in (6) as implemented in R package CGEN (6, 22, 25) directly into equation (3), namely, $\widehat{RERI}_{EB2} = \widehat{OR}_{GE} - \widehat{OR}_{G} - \widehat{OR}_{E} + 1$, where all estimated ORs are EB estimates proposed under the multiplicative model. The EB estimator we proposed in equation (7) demonstrates superior performance among the three choices, based on our simulation study.

Remark 2: As shown in Chen et.al (25), the asymptotic theory for CML and consequently EB is non-regular under the independence assumption. The Delta method does not technically apply for estimating 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 1-2 and Web Figures 1-2). Under G-E dependence, EB estimate converges in large sample to 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 Chatterjee and Carroll (22), Mukherjee et.al (6) and as implemented in the R package CGEN:

$$P(G,E,\boldsymbol{Z}|D) = \frac{P(D=1|G,E,\boldsymbol{Z})P(G|E,\boldsymbol{Z})P(E,\boldsymbol{Z})}{\sum_{G,E,\boldsymbol{Z}} P(D=1|G,E,\boldsymbol{Z})P(G|E,\boldsymbol{Z})P(E,\boldsymbol{Z})} \qquad (8)$$

The three ingredients of the above retrospective likelihood are:

(a) The logistic regression disease risk model of interest with multiplicative GEI parameter: $\text{logit}\,\mathrm{P}(D=1|G,E,\boldsymbol{Z}) = \beta_0 + \beta_G G + \beta_E E + \beta_{GE} G \times E + \boldsymbol{\beta}_{\boldsymbol{Z}}^{\boldsymbol{T}}\boldsymbol{Z}$, where $\boldsymbol{Z}$ denotes other covariates.

(b) $\text{logit}\,\mathrm{P}(G|E,\boldsymbol{Z}) = \theta_0 + \theta_{GE}E + \boldsymbol{\theta}_{\boldsymbol{GZ}}^{\boldsymbol{T}}\boldsymbol{Z}$. While this is the gene model used for UML, allowing G-E dependence, in the CML method, $\mathrm{P}(G|E,\boldsymbol{Z})$ reduces to $\mathrm{P}(G|\boldsymbol{Z})$ under the assumption of G-E independence conditional on $\boldsymbol{Z}$, implying $\theta_{GE} \equiv 0$.

(c) The distribution $\mathrm{P}(E,\boldsymbol{Z})$ is allowed to be completely non-parametric. We then maximize the retrospective likelihood using existing routines in CGEN to obtain $\widehat{\beta}_{uml}$ and $\widehat{\beta}_{cml}$, the vector of all the parameter estimates of the disease risk model in (a), namely, $(\beta_0, \beta_G, \beta_E, \beta_{GE}, \boldsymbol{\beta}_{\boldsymbol{Z}})$.

When it comes to defining RERI with a general *G* and *E* variable adjusting for covariates **Z**, particularly with case-control data, as described in VanderWeele (23), let us denote by $RERI_{OR}(E_0, E_1, G_0, G_1)$ the relative excess risk due to interaction by replacing risk ratios with corresponding odds ratios in the RERI expression in (3) as 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 when environmental risk factor changes from E_0 to E_1 and genetic risk factor changes from G_0 to G_1 but other covariates $\boldsymbol{z}$ are held constant. Formally, it is defined as

$$RERI_{OR}(E_0, E_1, G_0, G_1)$$
$$= OR(G_1, E_1) - OR(G_1, E_0) - OR(G_0, E_1) + 1$$

$$= \exp\{\beta_G(G_1 - G_0) + \beta_E(E_1 - E_0) + \beta_{GE}(G_1 \times E_1 - G_0 \times E_0)\}$$

$$- \exp\{\beta_E(E_1 - E_0) + \beta_{GE} G_0 \times (E_1 - E_0)\}$$

$$- \exp\{\beta_G(G_1 - G_0) + \beta_{GE}(G_1 - G_0) \times E_0\} + 1$$

$$= f(\beta_G, \beta_E, \beta_{GE}) \approx RERI(E_0, E_1, G_0, G_1) \quad (9)$$

This last approximation of risk ratios by odds ratios holds when the outcome is rare in each stratum defined by the two exposures or when controls are selected from the entire population, not just the non-cases (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 $\beta_G, \beta_E, \beta_{GE}$ in equation (9) become (I-1), (J-1) and (I-1)(J-1) dimensional vectors instead of scalars. Note that $\widehat{RERI}_{uml} = f(\widehat{\boldsymbol{\beta}}_{uml})$ and $\widehat{RERI}_{cml} = f(\widehat{\boldsymbol{\beta}}_{cml})$, can be viewed as function of UML and CML estimates of relative risk parameters, where f is the function in equation (9). The variance of $\widehat{RERI}_{uml}$ and $\widehat{RERI}_{cml}$ can be calculated by Delta method. The EB estimator of RERI is same as in equation (7) and its estimated variance is calculated by Delta method using the joint distribution of $(\widehat{\boldsymbol{\beta}}_{\boldsymbol{uml}}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{cml}})$ as proposed by Mukherjee et.al (6) (Web Appendix 6). The Wald tests for the three estimators are all based on the standardized Z statistic. We have provided general codes to test for RERI at (28).

Example: Analysis of G x 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 14,240 women died from ovarian cancer in 2016 in the United States, causing more deaths than any other cancer of the female reproductive system. There are several well-established non-genetic risk factors for ovarian cancer (29-35), and recent genome-wide association studies have identified and replicated 18 variants that influence disease risk (36). To this end, the Ovarian Cancer Association Consortium (OCAC) has undertaken an effort to study interactions focusing on the 18 confirmed single nucleotide polymorphisms (SNPs) and seven well-established risk factors: race, history of endometriosis, first degree family history of ovarian cancer, oral contraceptive pill (OCP) use, parity, tubal ligation, and age. In our illustrative analysis, we focus on OCP x SNP interaction and use genetic data from 15 OCAC studies that also have data on epidemiologic risk factors.

Each SNP is coded as the number of risk alleles a subject carried and all subsequent analysis assumed this additive genetic susceptibility model. Published ORs of the 18 confirmed loci in Web Table 3 are from analyses presented in Collaborative Oncological Gene-Environment Study (37-44). 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 OR in large meta-analysis. Polygenic risk scores have been used for risk stratification in multiple G x E papers recently (3,45). Analysis of marginal effect for GRS and WGRS is shown in Web Table 4. Each environmental factor is coded as a categorical variable as described in Web Table 5. The merged G × E dataset has a sample size of 11,661 subjects with European ancestry, with 4,135 cases and 7,526 controls from 13 study sites (Web Table 6).

To illustrate our inference for interactions between OCP use (1 =ever and 0 =never) and genetic risk factors we consider both single SNP x OCP and (W)GRS x OCP interaction. For single SNP analysis, we consider the top two hits in the 18 confirmed loci, i.e. rs62274042 (SNP1) and rs10962691 (SNP2) as reported in Web Table 3. We used additive coding for our SNP x OCP analysis. For GRS and WGRS, we use the quartiles in controls to define a categorical variable with four 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 $\widehat{RERI}_{uml}, \widehat{RERI}_{cml}$ and $\widehat{RERI}_{EB}$ under various alternative values of RERI across a spectrum of scenarios, varying the strength of G-E association, main effects of G and E, minor allele frequency of G, prevalence of exposure E, test size and sample sizes. We compare the power of Wald test for $\widehat{RERI}_{cml}$ 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 three estimators as well as those of two alternative proposals, $\widehat{RERI}_{EB1}$ and $\widehat{RERI}_{EB2}$. Note that both RERI and multiplicative interaction parameters are obtained from the underlying true logistic regression model

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

where 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 $OR_G = \exp(\beta_G)$ and $OR_E = \exp(\beta_E)$ are specified.

We set prevalence 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 $OR_G = (1.1, 1.2, 1.3)$; $OR_E = (1.3, 1.5, 1.7)$; sample size $N_0 = N_1 = (4000, 20000)$;

size of test $\alpha = (0.05, 5 \times 10^{-6})$; the strength of G-E association, $\exp(\theta_{GE})$, change from 0.8 to 1.2 at a grid of 0.1 and RERI change from 0 to 1.5 with a grid of 0.1. The number of simulated datasets is 1000 when $\alpha = 0.05$ and 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, 46): We generate data independently from the two multinomial distributions corresponding to the case and control populations, according to the above probabilities with number of cases and control as $N_0,\ N_1$, respectively. We also considered another simulation setting to mimic a large-scale genomewide 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 and MSE. Type I error over 1000 replications. Power are estimated by the proportion of null hypothesis $\mathrm{H}_{\mathbf{0}}: RERI = 0$ rejected at the given level of significance α, i.e. the proportion of times $|Z|$>$Z_{1-\alpha/2}$, where Z is 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 $\left|\widehat{RERI} - RERI\right|/RERI$ and MSE is calculated by averaging $\left(\widehat{RERI} - RERI\right)^2$.

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 in both multiplicative and additive scale. The estimates, corresponding CIs and P-values of Wald test are shown in Table 1. In SNP1×OCP analysis, the strength of G-E association is modest: $\exp(\theta_{\mathrm{GE}})$=1.07 (95% CI [0.94,1.21]), EB estimate of RERI is -0.16 with 95% CI [-0.50,0.18], where the weight on $\widehat{RERI}_{uml}$ is 43%. In SNP2×OCP analysis, the G-E association seems weaker with $\exp(\theta_{\mathrm{GE}})$=0.96 (95% CI [0.83,1.11]). EB estimate of RERI is 0.04 with 95% CI [-0.11,0.18], with its weight on $\widehat{RERI}_{uml}$ decreasing to 11%. The confidence intervals corresponding to $\widehat{RERI}_{EB}$ are narrower compared to the corresponding intervals for $\widehat{RERI}_{uml}$. The point estimate $\widehat{RERI}_{EB}$ lies between $\widehat{RERI}_{uml}$ and $\widehat{RERI}_{cml}$, reflecting the combined efficiency-robustness feature of the EB estimator. In WGRS×OCP analysis we report interactions associated with a change of OCP from 0 to 1 (ever users to never users) and WGRS from the lowest to the highest quartile (as defined through distribution of WGRS in controls) the multiplicative measure of interaction $\hat{\psi}_{EB}$ is not significant at α=0.05 but $\widehat{RERI}_{EB}$ departs from 0 significantly with EB estimate of RERI - 0.52(95% CI [-0.91, -0.13]) and has a very small P-value, 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 of OCP and corresponding CI stratified by WGRS. The odds ratio of OCP is 0.61 (0.50,0.74) in the lowest WGRS quartile and 0.51 (0.43,0.60) in the highest quartile. The overlapping CIs indicate a non-significant multiplicative interaction. Additionally, if we assume that approximately 1.3 percent of women will be diagnosed with ovarian cancer at some point during their lifetime (47) and 70% women will use OCP at some point in their life in this population (estimated from the OCAC data), we present the estimated lifetime risk of ovarian cancer and corresponding 95% CI within each WGRS stratum in Figure 2, for OCP users and non-users. Estimates of lifetime absolute risk for OCP users is 0.75% (0.57%, 0.98%) and 1.23% (1.00%, 1.51%) for OCP non-users in the lowest WGRS stratum with a difference of 0.48% (0.02%, 0.94%) and the corresponding numbers were 1.40% (1.08%, 1.81%) and 2.72% (2.05%, 3.60%) with a difference of 1.32% (0.24%, 2.52%) for subjects in the highest WGRS stratum, 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 nominal level α across different choices of θ_{GE}. An inflated Type I error associated with CML is observed when G-E independence assumption is violated. EB test is valid when $\exp(\theta_{\mathrm{GE}})$=1 and has a modest inflation on Type I error when G is associated with E. The maximal observed Type I error of EB at α=0.05 is 0.099 when sample size is 40,000, test size is 0.05 and $\exp(\theta_{\mathrm{GE}})$=1.1. Web Figure 4 presents how Type I error varies with $\exp(\theta_{\mathrm{GE}})$ for the three estimators. The Type I error of CML is very sensitive to the G-E association but the performance of EB is relatively robust with marked reduction in Type I error compared to CML. The findings remain similar for different choices of p_G, p_E, OR_G and OR_E (Web Tables 8-9).

Results from additional simulation mimicking a Genomewide Association Study: To justify the use of EB estimator in genomewide assessment of G-E interaction, we conduct another simulation study similar to that in Reference (8), which generates 2000 cases and controls with 1 causal marker together with M-1 null markers where M is 10,000. G-E independence parameter θ_{GE} in controls have a random mixture distribution with point mass around independence and $\mathrm{p}_{\mathrm{ind}}$ 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 familywise error rate and expected number of false positives 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 familywise Type I error rate of 0.06 and expected number of false positives of 0.06. The performance of CML is significantly worse with familywise error rate of 99% and expected number of false positives 3.76.

Power. Figure 3 shows the power curves of Wald test for three estimators with $\mathrm{H}_0: RERI = 0$ 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 the misleading 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 the most efficient when exp(θ_{GE})=0.8, CML is the most efficient when exp(θ_{GE})=1 and 1.2, and EB power always lies in between. For a sample numerical comparison, let us compare the powers of the three approaches at RERI=0.5 to represent one typical scenario. When exp(θ_{GE})=0.8, the empirical power of EB (0.275) is 41% lower than UML (0.672), meanwhile CML has nearly 0 power. When exp(θ_{GE})=1, the empirical power of EB (0.870) is 25% higher than UML (0.693) but 10% lower than CML (0.970). When exp(θ_{GE})=1.2, the empirical power of EB (0.718) is slightly higher than UML (0.714) but 28% lower than CML (0.993). We then compare the power of Wald test for $\widehat{RERI}_{cml}$ with LRT for additive interaction shown in Web Figure 5. The power of LRT is uniformly slightly higher than the Wald test with true value of RERI varying from 0 to 0.5 with a grid of 0.1.

Absolute relative bias and MSE results are relegated to Web Appendix 9, Web Tables 16-19, Web Figure 6.

DISCUSSION

In this paper, we extend the EB estimator of gene-environment interaction proposed earlier on the multiplicative scale to additive scale in case-control studies. The EB estimator exploits G-E independence assumption to perform a trade-off between bias and efficiency. The simulation study showed that the test based on the EB estimator can provide a good control of Type I error and 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 α=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. $\mathrm{p}_{\mathrm{additive}}$ 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 (48), as we carried out in Figure 2. If the rare disease assumption for each configuration of G and E does not hold, approximating RR by OR 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, using prior guesses for disease prevalence and adaptive combinations like EB procedure we can make our inference less biased under violation of the rare disease and gene-environment independence assumptions.

There is increasingly more interest in inference for additive interaction using case-control data. Tchetgen –Tchetgen et.al (49) described a general approach to test for G x E additive interaction exploiting G-E independence which is robust to possible misspecification of main effects in the outcome regression. Han et.al (50) proposed a score test for UML and CML estimators of genetic associations under the additive null. In the future, it is of analytical interest to establish an EB version of adaptive score test and adaptive LRT as most of the recent work has been in terms of combining estimators but not tests.

ACKNOWLEDGEMENTS

Author affiliations: Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA (Gang Liu, Seunggeun Lee, and Bhramar Mukherjee); Department of Preventive Medicine, University of Southern California Keck School of Medicine, Los Angeles, California, USA (Alice W. Lee, Anna H. Wu, Malcolm C. Pike, Celeste Leigh Pearce); Department of Cancer Epidemiology, Division of Population Sciences, Moffitt Cancer Center, Tampa, Florida, USA (Catherine M. Phelan); Cancer Prevention and Control, Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey, USA (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, USA (Mary Rossing); Department of Epidemiology, University of

Washington, Seattle, Washington, USA (Mary Rossing); Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, New York, USA (Kirsten B. Moysich); Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany (Jennifer Chang-Claude); University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany (Jennifer Chang-Claude); Department of Epidemiology, The Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA (Jennifer Doherty); Women's Cancer, Institute for Women's Health, University College London, London, United Kingdom (Aleksandra Gentry-Maharaj); 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, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA (Francesmary Modugno); Department of Epidemiology, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania, USA (Francesmary Modugno); Womens Cancer Research Program, Magee-Womens Research Institute and University of Pittsburgh Cancer Institute, Pittsburgh, Pennsylvania, USA (Francesmary Modugno); Radboud University Medical Center, Radboud Institute for Molecular Life Sciences, Nijmegen, The Netherlands (Leon Massuger); Department of Health Sciences Research, Division of Epidemiology, Mayo Clinic, Rochester, Minnesota, USA (Ellen L. Goode); Kansas IDeA Network of Biomedical Research Excellence Bioinformatics Core, University of Kansas Cancer Center, Kansas City, Kansas, USA (Brooke Fridley); Obstetrics and Gynecology Center, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA (Kathryn L. Terry, Daniel W. Cramer); Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA (Kathryn L. Terry, Daniel W. Cramer);

Genetic Epidemiology Research Institute, UCI Center for Cancer Genetics Research and Prevention, School of Medicine, Department of Epidemiology, University of California Irvine, Irvine, California, USA (Hoda Anton-Culver, Argyrios Ziogas); Department of Public Health and Primary Care, Center for Cancer Genetic Epidemiology, University of Cambridge, Strangeways Research Laboratory, Cambridge, UK (Jonathan Tyrer, Paul D. Pharoah); Department of Public Health Sciences, The University of Virginia, Charlottesville, Virginia, USA (Joellen M. Schildkraut); Department of Gynecology, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark (Susanne K. Kjaer); Queensland Institute of Medical Research, Brisbane, Australia (Penelope M. Webb); University of Texas School of Public Health, Houston, Texas, USA (Roberta B. Ness); Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York, USA (Malcolm C. Pike); Department of Women's Cancer, EGA Institute for Women's Health, University College London, London, United Kingdom (Usha Menon); Department of Obstetrics and Gynecology, Duke University Medical Center, Durham, North Carolina, USA (Andrew Berchuck); Department of Oncology, Center for Cancer Genetic Epidemiology, University of Cambridge, University of Cambridge, Cambridge, United Kingdom (Paul D. Pharoah); Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, Connecticut, USA (Harvey Risch); and Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, Michigan, USA (Celeste Leigh Pearce).

This work was supported by: the US National Cancer Institute (R01 CA076016); the COGS project is funded through a European Commission's Seventh Framework Programme grant (agreement number 223175 HEALTH F2 2009-223175); the Genetic Associations and Mechanisms in Oncology (GAME-ON): a NCI Cancer Post-GWAS Initiative (U19-CA148112); 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 (PPD/RPCI.07).

It was also supported by the National Institutes of Health (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]); California Cancer Research Program (0001389V20170 and 2110200 [USC]); German Federal Ministry of Education and Research of Germany, Programme of Clinical Biomedical Research (01GB9401 [GER]); German Cancer Research Centre (GER); Danish Cancer Society (94 222 52 [MAL]); Mermaid I [MAL]; Eve Appeal/Oak Foundation (UKO); Cancer Institute of New Jersey (NJO); the National Institute for Health Research University College London Hospitals Biomedical Research Centre (UKO); US Army Medical Research and Materiel Command (W81XWH-10-1-02802 [NEC], DAMD17-02-1-0669 [HOP], DAMD17-02-1-0666 [NCO], and DAMD17-01-1-0729 [AUS]); Roswell Park Alliance Foundation (HOP); Cancer Councils of New South Wales, Victoria, Queensland, South Australia and Tasmania (Multi-State Application Numbers 191, 211 and 182 [AUS]); Cancer Foundation of Western Australia (AUS); National Health and Medical Research Council of Australia (199600 and 400281 [AUS]); Mayo Foundation (MAY); Minnesota Ovarian Cancer Alliance (MAY); Fred C. and Katherine B. Andersen Foundation (MAY); Radboud University

Medical Centre (NTH); Lon V Smith Foundation (LVS-39420 [UCI]); National Institute of Environmental Health Sciences (T32 ES013678 for A.W.L.); National Health and Medical Research Council of Australia (for G.C.T. and P.W.).

Research reported in this publication was partially supported by NCI award number P30 CA008748 (PI: Thompson) to Memorial Sloan Kettering Cancer Center. It was also supported by the National Cancer Institute of the National Institutes of Health (P30 CA046592). Lastly, this work was also supported by the National Science Foundation (NSF DMS 1406712) and the National Institutes of Health (NIH ES 20811). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

We thank all the individuals who took part in this study and all 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 (Genome Quebec); Dr. U. Eilber (GER); Dr. I. Jacobs, Dr. M. Widschwendter, Dr. E. Wozniak, N. Balogun, Dr. A. Ryan and J. Ford (UKO).

Conflict of interest: None declared.

REFERENCES

1. Hutter C, Chang-Claude J, Slattery M, et al. Characterization of Gene-Environment Interactions for Colorectal Cancer Susceptibility Loci. *Cancer Research*. 2012; 72(8): 2036-2044.
2. Hsu L, Jiao S, Dai J, et al. Powerful Cocktail Methods for Detecting Genome-Wide Gene-Environment Interaction. *Genetic Epidemiology*. 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 Research*. 2013; 73(7): 2211-2220.
4. Figueiredo J, Hsu L, Hutter C, et al. Genome-Wide Diet-Gene Interaction Analyses for Risk of Colorectal Cancer. *PLoS Genetics*. 2014; 10(4): p.e1004228.
5. Lewinger J, Morrison J, Thomas D, et al. Efficient Two-Step Testing of Gene-Gene Interactions in Genome-Wide Association Studies. *Genetic Epidemiology*. 2013; 37(5): 440-451.
6. Mukherjee B and 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 C, Lewinger, J and Gauderman W. Gene-Environment Interaction in Genome-Wide Association Studies. *American Journal of Epidemiology*. 2008; 169(2): 219-226.
8. Mukherjee B, Ahn J, Chatterjee N, et al. Testing Gene-Environment Interaction in Large-Scale Case-Control Association Studies: Possible Choices and Comparisons*. American Journal of Epidemiology,* 2012; 175(3): 177-190.

9. Boonstra P, Mukherjee B, Chatterjee N, et al. Tests for Gene-Environment Interactions and Joint Effects with Exposure Misclassification. *American Journal of Epidemiology*. 2016; 183(3): 237-247.

10. Thomas D. Methods for Investigating Gene-Environment Interactions in Candidate Pathway and Genome-Wide Association Studies. *Annual Review of Public Health*. 2010; 31(1): 21-36.

11. Prentice R and Pyke R. Logistic Disease Incidence Models and Case-Control Studies. *Biometrika*.1979; 66(3): 403.

12. Piegorsch W, Weinberg C and Taylor J. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statist. Med.*1994; 13(2): 153-162.

13. Umbach D and Weinberg C. Designing and analyzing case-control studies to exploit independence of genotype and exposure. *Statist. 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 Epidemiology Biomarkers & Prevention*. 2014; 23(12): 2971-2976.

15. Joshi A, Lindstrom S, Husing 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. *American Journal of Epidemiology*. 2014; 180(10): 1018-1027.

16. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Philadelphia: Wolters Kluwer Health; 2015: 71-87.

17. VanderWeele T. A Word and That to Which it Once Referred. *Epidemiology*. 2011; 22(4): 612-613.

18. Hosmer D and 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. *American Journal of Epidemiology*. 2008; 168(2): 212-224.

20. VanderWeele T. Sample Size and Power Calculations for Additive Interactions. *Epidemiologic Methods*. 2012; 1(1): 159-188.

21. Han S, Rosenberg P, 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. *American Journal of Epidemiology*. 2012; 176(11): 1060-1067.

22. Chatterjee N and Carroll R. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*. 2005; 92(2): 399-418.

23. VanderWeele TJ. 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. *Statist. Med*. 2008; 27(15): 2756-2783.

25. Chen YH, Chatterjee N and Carroll R. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of the American Statistical Association.* 2009; 104: 220-233.

26. Leeb H and Pötscher BM. Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics.* 2008;142(1):201–211.

27. Knol M, Vandenbroucke J, Scott P, et al. What Do Case-Control Studies Estimate? Survey of Methods and Assumptions in Published Case-Control Research. *American Journal of Epidemiology*. 2008; 168(9): 1073-1081.

28. Github website. https://github.com/GreysonL/RERI/releases. Updated January 12, 2017. Accessed May 12, 2017.

29. Beral V, Doll R, Hermon C, 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.

30. 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.

31. 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.

32. Tung KH, Goodman MT, Wu AH, et al. Reproductive factors and epithelial ovarian cancer risk by histologic type: a multiethnic case-control study. *American Journal of Epidemiology*. 2003; 158(7):629–638.

33. Cibula D, Widschwendter M, Majek O, et al. Tubal ligation and the risk of ovarian cancer: review and meta-analysis. *Hum Reprod Update*. 2011; 17(1):55–67.

34. 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.

35. Auranen A, Pukkala E, Makinen J, et al. Cancer incidence in the first- degree relatives of ovarian cancer patients. *Br J Cancer*. 1996; 74(2):280–284.

36. Pearce CL, Rossing M, Lee, et al. Combined and Interactive Effects of Environmental and GWAS-Identified Risk Factors in Ovarian Cancer. *Cancer Epidemiology Biomarkers & Prevention.* 2013; 22(5): 880-890.

37. Kuchenbaecker, et al. Identification of six new susceptibility loci for invasive epithelial ovarian cancer. *Nat Genet*. 2015; 47(2): 164-171.

38. Bolton KL, et al. Common variants at 19p13 are associated with susceptibility to ovarian cancer. *Nat Genet*. 2010; 42:880–884.

39. Goode EL, et al. A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nat Genet*. 2010; 42:874–879.

40. Song H, et al. A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2. *Nat Genet*. 2009; 41:996–1000.

41. Pharoah PD, et al. GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nat Genet.* 2013; 45:362–370. 370e361–370e362.

42. Permuth-Wey J, et al. Identification and molecular characterization of a new ovarian cancer susceptibility locus at 17q21.31. *Nat Commun*. 2013; 4:1627.

43. Bojesen 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; 45:371–384. 384e371–384e372.

44. Couch FJ, et al. Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLoS Genet*. 2013; 9:e1003212.

45. Li S, Zhao J, Loos RJ, 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.

46. Mukherjee B, Ahn J, Chatterjee N, et al. Tests for gene-environment interaction from case-control data: a novel study of Type I error, power and designs. *Genetic Epidemiology*. 2008; 32(7): 615-626.

47. National Cancer Institute, Surveillance, Epidemiology, and End Results Program, SEER Stat Fact Sheets—Ovarian Cancer. http://seer.cancer.gov/statfacts/html/ovary.html. Published April 14, 2017. Accessed May 12, 2017.

48. Risch HA, Yu H, Lu L, et al. Detectable symptomatology preceding the diagnosis of pancreatic cancer and absolute risk of pancreatic cancer diagnosis. *American Journal of Epidemiology*. 2015; 182(1): 26-34.

49. Tchetgen E, Sofer T and Wong B. A General Approach to Detect Gene (G)-environment (E) Additive Interaction Leveraging G-E Independence in Case-control Studies. [online] Collection of Biostatistics Research Archive. Available at: http://biostats.bepress.com/harvardbiostat/paper177/ [Accessed 30 Jun. 2014].

50. Han S, Rosenberg P, Chatterjee N, et al. An exposure-weighted score test for genetic associations integrating environmental risk factors. *Biometrics*. 2015; 71(3): 596-605.

Table 1. Estimates and 95% confidence interval corresponding to SNP/GRS x Oral Contraceptive Pill Use Interaction under Both Multiplicative and Additive Scale with accompanying P-values from Wald Tests

Interaction	**Multiplicative (ψ)**			**Additive (RERI)**		
SNP1[a]×OCP[b]	Estimate[c]	95% CI	P-value	Estimate[d]	95% CI	P-value
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[a]×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[d]×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[d]×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: CML, constrained maximum-likelihood; EB, empirical Bayes; GRS, genetic risk score; RERI, relative excess risk due to interaction; UML, unconstrained maximum-likelihood; WGRS, weighted genetic risk score.

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

[b] OCP=1 if the individual ever used OCP and OCP=0 if never.

[c] The analysis is based on subjects with European ancestry, using data on 4,135 cases and 7,526 controls from 13 study sites from the Ovarian Cancer Association Consortium. The model adjusts for history of endometriosis, first degree family history of ovarian cancer, parity, tubal ligation, age and study site.

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

Table 2. Empirical Familywise Type I Error Rate at 5% overall level of significance, and Expected Number of False Positives corresponding to UML, CML and EB Wald Tests

	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 number 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; UML, unconstrained maximum-likelihood.

[a] The population-level G-E association structure among null loci is assumed to be of the form of a mixture distribution reflecting that a large fraction, i.e., $\mathrm{p_{ind}}$, of the SNPs, indeed, are independent of E in the population, whereas the remaining $(1 - p_{ind})$ of SNPs show some departures from the independence assumption following a N (0, sd=log(1.5)/2) distribution.

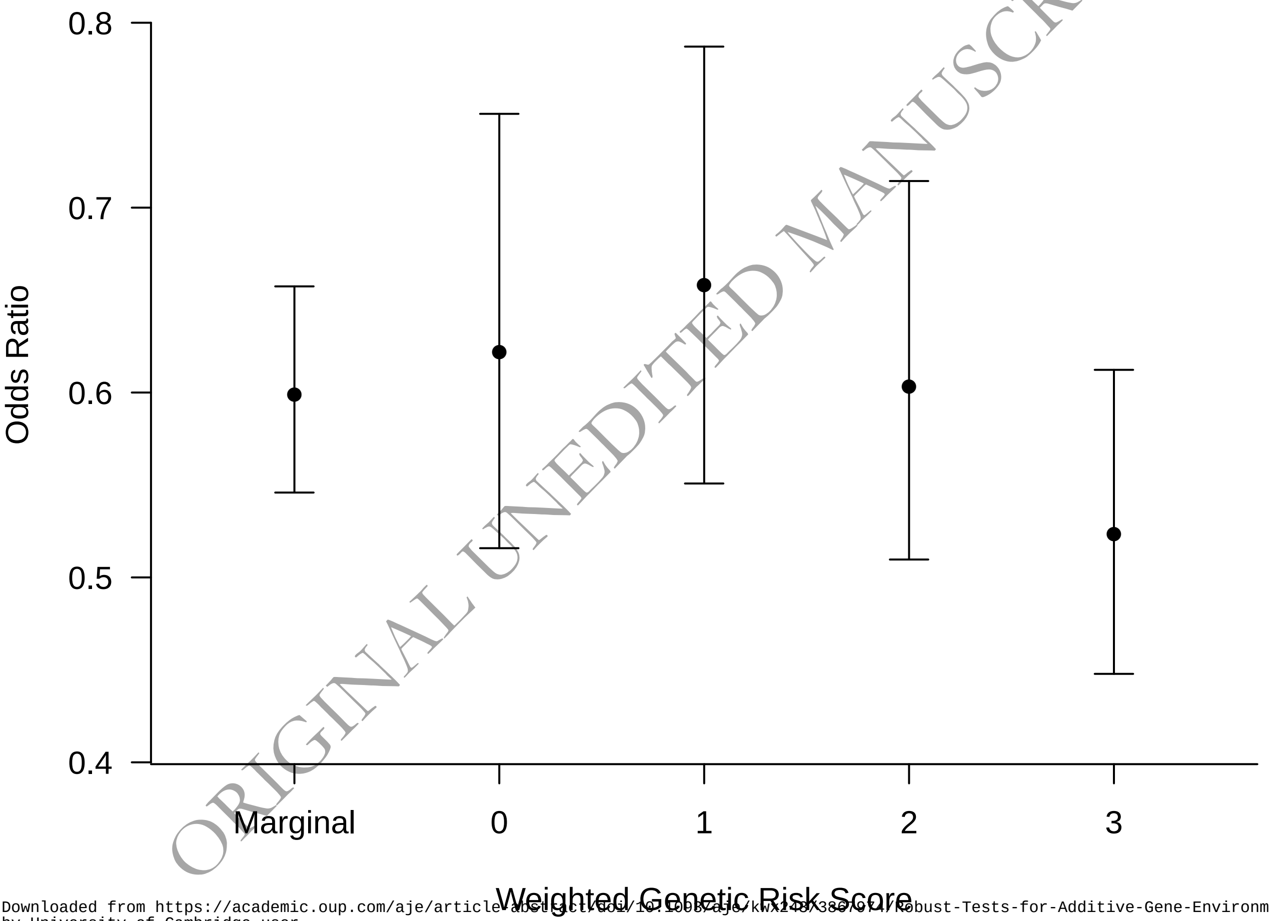
[b] The Wald test is for RERI=0 under a large-scale genomewide G x E scan simulation scenario with 10000 markers and 2000 cases and controls. Empirical familywise type I error is estimated as the empirical proportion of data sets declaring at least 1 null marker to be significant using level of significance α/10000. This estimates the probability of at least one false positive under the global null.

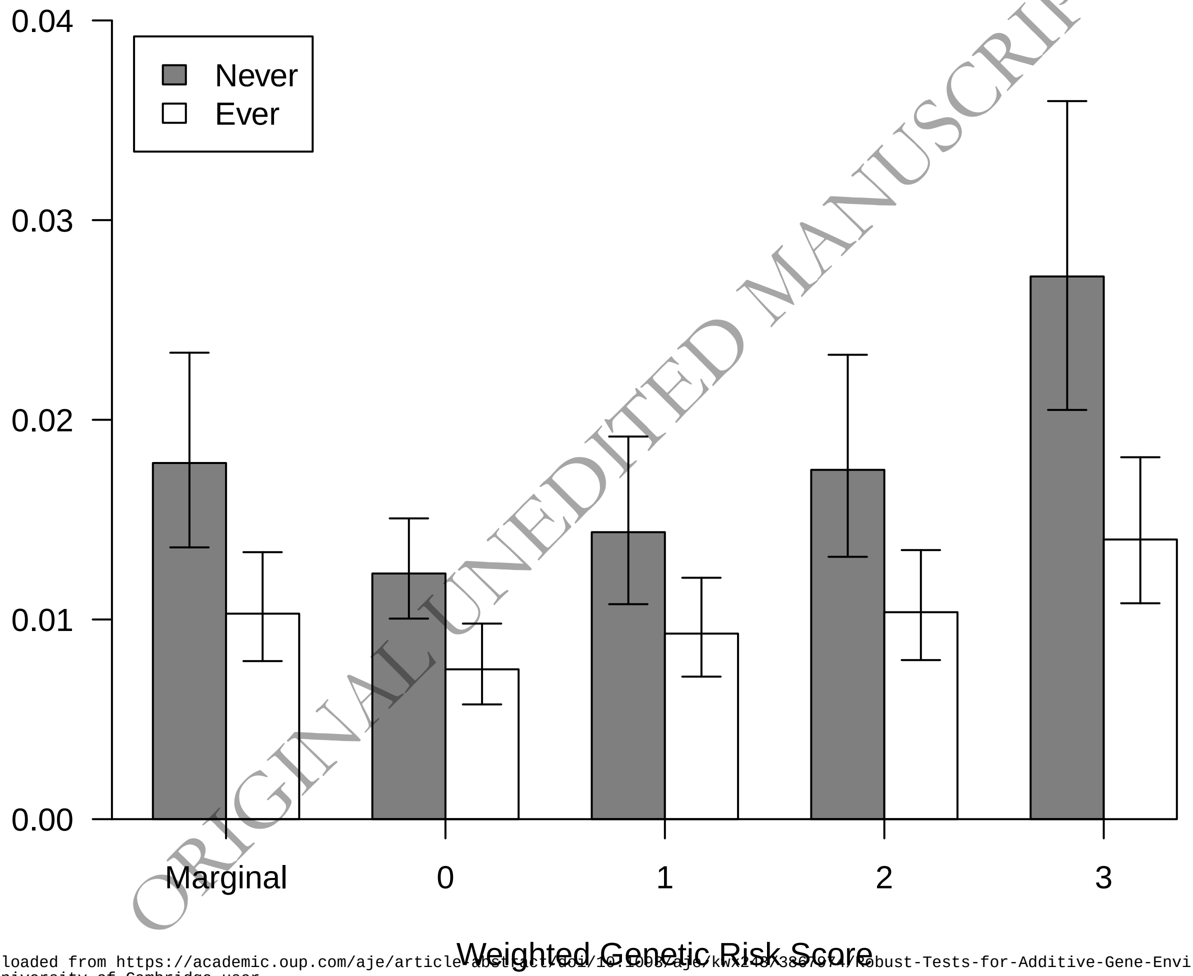
[c] Expected number of false positives is estimated as the average number of falsely rejected null hypotheses, averaged over 1000 data sets.

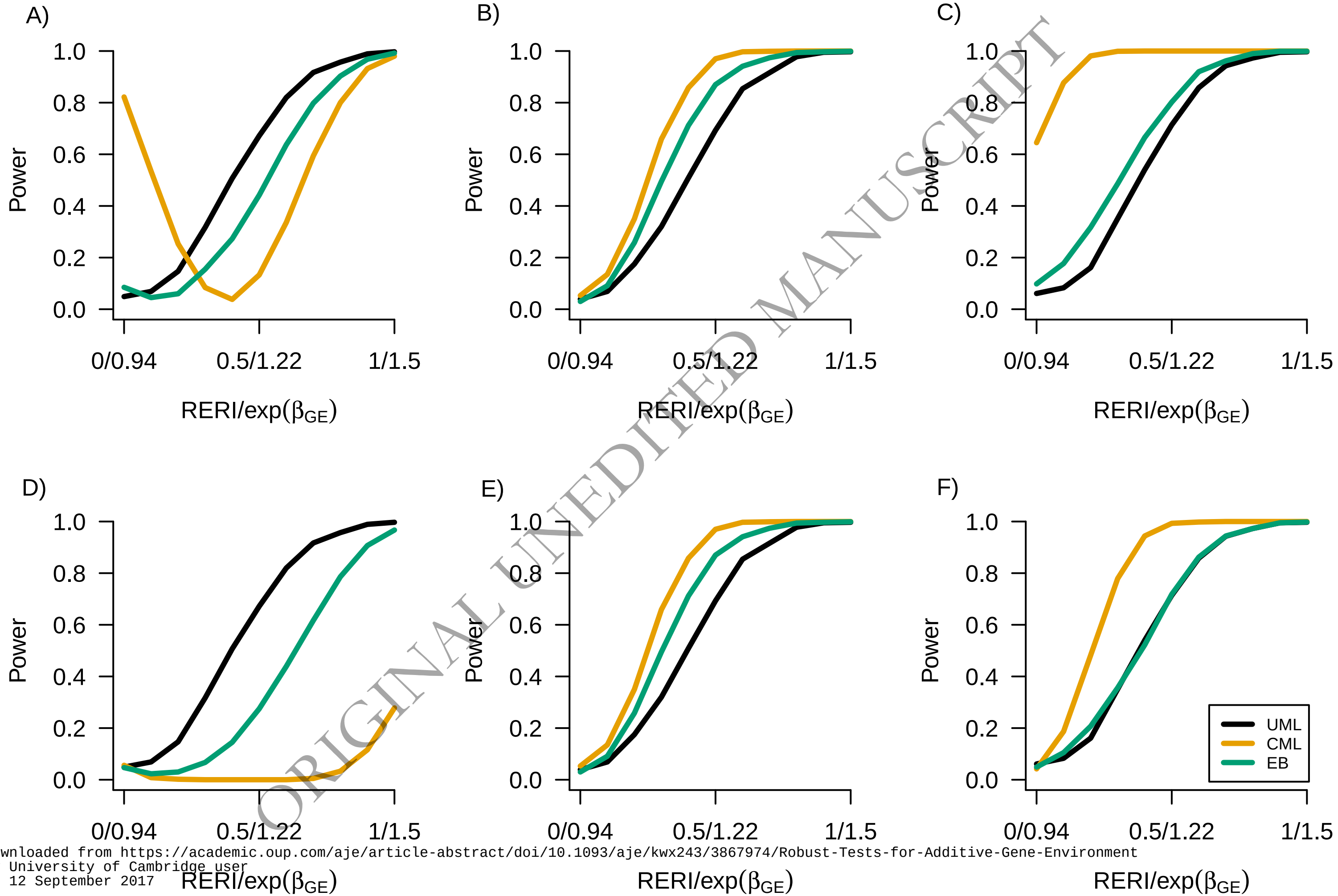
Figure 1. Odds ratio of oral contraceptive pill and corresponding 95% CI within each quartile of the weighted genetic risk score. The odds ratios are estimated from a standard logistic regression adjusting for history of endometriosis, first degree family history of ovarian cancer, parity, tubal ligation, age and study site.

Figure 2. Predicted probability of ovarian cancer and corresponding 95% CI within each quartile of the weighted genetic risk score comparing oral contraceptive pill users and non-users. The relative risk parameters are 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 assume that approximately 1.3 percent of women will be diagnosed with ovarian cancer at some point during their lifetime and 70% women will use oral contraceptive pill at some point in their life. The predicted probabilities are estimated by fixing other covariates at their most frequent value.

Figure 3. Power curves of unconstrained maximum-likelihood (UML), constrained maximum-likelihood (CML) and empirical Bayes (EB) Wald test for relative excess risk due to interaction (RERI) under different strength of G-E association: data are generated on 4000 cases and 4000 controls with fixed parameters $p_G = 0.2$, $p_E = 0.3$, $OR_G = 1.2$, $OR_E = 1.5$. RERI changes from 0 to 1.5 with a grid level of 0.1, corresponding multiplicative interaction changes from 0.94 to 1.78. The top panels (A, B, C) correspond to the raw power, whereas the bottom panels (D, E, 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, i.e. exp(θ_{GE})=0.8, 1.0, 1.2.


Odds Ratio
0.8
0.7
0.6
0.5
0.4
Marginal
0
1
2
3
Weighted Genetic Risk Score


Never
Ever
0.04
0.03
0.02
0.01
0.00
Marginal
0
1
2
3
Weighted Genetic Risk Score


A)
B)
C)
D)
E)
F)
1.0
0.8
0.6
0.4
0.2
0.0
Power
0/0.94
0.5/1.22
1/1.5
RERI/exp(β_{GE})
UML
CML
EB