Tests for Gene-Environment Interaction From Case-Control Data: A Novel Study of Type I Error, Power and Designs

Bhramar Mukherjee,[1] Jaeil Ahn,[1] Stephen B. Gruber,[2] Gad Rennert,[3] Victor Moreno[4] and Nilanjan Chatterjee[5]*

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan
[2]Department of Internal Medicine, Epidemiology and Human Genetics, University of Michigan, Ann Arbor, Michigan
[3]Department of Community Medicine and Epidemiology, Carmel Medical Center and Technion Faculty of Medicine, CHS National Israeli Cancer Control Center, Haifa, Israel
[4]IDIBELL, Catalan Institute of Oncology, L'Hospitalet, Barcelona, Spain
[5]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Rockville, Maryland

To evaluate the risk of a disease associated with the joint effects of genetic susceptibility and environmental exposures, epidemiologic researchers often test for non-multiplicative gene-environment effects from case-control studies. In this article, we present a comparative study of four alternative tests for interactions: (i) the standard case-control method; (ii) the case-only method, which requires an assumption of gene-environment independence for the underlying population; (iii) a two-step method that decides between the case-only and case-control estimators depending on a statistical test for the gene-environment independence assumption and (iv) a novel empirical-Bayes (EB) method that combines the case-control and case-only estimators depending on the sample size and strength of the gene-environment association in the data. We evaluate the methods in terms of integrated Type I error and power, averaged with respect to varying scenarios for gene-environment association that are likely to appear in practice. These unique studies suggest that the novel EB procedure overall is a promising approach for detection of gene-environment interactions from case-control studies. In particular, the EB procedure, unlike the case-only or two-step methods, can closely maintain a desired Type I error under realistic scenarios of gene-environment dependence and yet can be substantially more powerful than the traditional case-control analysis when the gene-environment independence assumption is satisfied, exactly or approximately. Our studies also reveal potential utility of some non-traditional case-control designs that samples controls at a smaller rate than the cases. Apart from the simulation studies, we also illustrate the different methods by analyzing interactions of two commonly studied genes, *N*-acetyl transferase type 2 and glutathione *s*-transferase M1, with smoking and dietary exposures, in a large case-control study of colorectal cancer. *Genet. Epidemiol*. 32:615–626, 2008. Published 2008 Wiley-Liss, Inc.[†]

Key words: case-only designs; empirical Bayes; gene-environment interaction; genome-wide scan; Molecular Epidemiology of Colorectal Cancer

Additional Supporting Information may be found in the online version of this article.
Contract grant sponsor: NSF; Contract grant number: DMS 07-06935; Contract grant sponsor: NIH; Contract grant numbers: R03 CA130045-01, R01 CA81488; Contract grant sponsor: Spanish Secretaria de Estado de Universidades e Investigacion; Contract grant number: PR2006-04743; Contract grant sponsor: National Heart Lung and Blood Institute; Contract grant number: R01 HL091172-01; Contract grant sponsor: National Cancer Institute.

*Correspondence to: Nilanjan Chatterjee, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Rockville, Maryland. E-mail: chattern@mail.nih.gov

Received 1 December 2007; Revised 13 February 2008; Accepted 5 March 2008
Published online 12 May 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20337

INTRODUCTION

The completion of the Human Genome Project and rapid advancement of genotyping technologies now hold great promise for discovering the inherited causes of a complex disease by studying genetic variations across candidate genes, biochemical pathways and the whole genome. To realize the full potential of these advances, however, it is critical to recognize that most common human diseases have a multifactorial etiology involving complex interplay of genetic susceptibility and environmental exposures and studying these factors together can improve the statistical power for detection of the underlying risk factors, give insight into their biologic effects and lead to public health strategies for prevention. An important step for characterization of gene-environment joint effects involves evaluation of statistical interaction or effect modification; i.e. whether the effect of one exposure, measured in a suitable scale, varies by the level of the other and vice versa.

Population-based case-control studies are now commonly used to study the roles of genes and gene-environment interactions in determining the risk of complex diseases. The goal of this article is to evaluate performances of alternative analytic methods and study designs for testing of non-multiplicative gene-environment effects from case-control studies. It is well known that standard case-control analysis often has poor power for detection of multiplicative interaction due to small numbers of cases or controls in cells of crossing genotypes and exposures. In contrast, under the assumption of gene-environment (*G-E*) independence for the underlying population, one can test for multiplicative interaction in a very powerful fashion based on the genotype-exposure odds-ratio among the cases alone [Piegorsch et al., 1994], but the method can have seriously inflated Type I error when the underlying assumption of gene-environment independence is violated [Albert et al., 2001]. To resolve the bias vs. efficiency dilemma, practitioners may naturally adapt to a two-stage procedure where, at first, one formally tests for the adequacy of the gene-environment independence assumption based on the data itself and then uses the outcome of that test to decide whether to use the powerful case-only or the more robust case-control test for estimation. For a given study of modest sample size, however, the power of the tests for gene-environment independence would be typically low and consequently the two-stage procedure, as a whole, could still remain significantly biased. Moreover, a proper variance calculation for the two-stage estimator accounting for the underlying model uncertainty can be fairly complicated. The standard two-stage testing procedure that ignores this model uncertainty maintains a higher Type I-error level than desired even when the gene-environment independence assumption is satisfied [Albert et al., 2001].

We have recently proposed a solution to the bias vs. efficiency dilemma by considering a novel composite estimator of the multiplicative interaction parameter obtained by taking a simple weighted average of the case-control and case-only estimators [Mukherjee and Chatterjee, 2007; referred to as MC from here onwards]. The weights are constructed in a data-adaptive way, so that in large sample the estimator has no bias irrespective of whether the *G-E* independence assumption holds or not and yet it can gain efficiency over the standard case-control estimator when the independence assumption is satisfied, exactly or approximately. We have shown that the proposed method, termed as empirical-Bayes (EB)-type shrinkage estimator, can achieve balance between bias and efficiency in the sense that it maintains optimal or close to optimal mean-squared-errors among all of the different estimators of interactions irrespective of the true state of the gene-environment association.

The purpose of this article is to provide a comparative study of alternative estimators of gene-environment interactions in terms of Type I error and power of the corresponding testing procedure. Our study has two unique aspects. We consider novel case-control designs where controls are sampled at a smaller ratio than the cases. Traditional power calculations suggest that power for a case-control study with 1:m case-control sampling ratio depends on m, the number of controls per case, through the ratio $m/m+1$ [Breslow and Day, 1987, pp289–294]. Such calculations suggest that while sampling controls at a much higher ratio than the cases (e.g. $m>4$) is not very beneficial, sampling them at a lower rate than the cases (e.g. $m=0.5$ or 0.25) can seriously hurt efficiency. Thus, case-control studies generally sample controls at least with an equal rate as the cases with the value of m typically ranging between 1 and 4. In contrast, if one could rely on the gene-environment independence assumption for the underlying population, then one could perform the test of interaction based on the case-only design, without sampling any controls at all. In this article, we consider intermediate study designs where controls are sampled to protect against inflated Type I error of the case-only approach when the *G-E* independence assumption is violated; but unlike traditional settings we consider case:control sampling ratio of 1:m with $m\leq 1$. The performance of the novel EB estimator both in terms of Type I error and power in such settings reveals interesting design considerations for future case-control studies.

Researchers traditionally evaluate performance of frequentist statistical methods under a fixed set of values for the underlying unknown parameters. A key parameter that determines the performance of alternative gene-environment interaction test is the log-odds-ratio between genotype-exposure in the underlying population, say denoted by θ_{GE}. When $\theta_{GE}=0$, i.e. the gene-environment independence assumption is satisfied, the case-only test maintains Type I error at the nominal level and has the highest power among all of the different tests for detecting interactions. When θ_{GE} departs from zero, on the other hand, the case-only estimator will perform poorly in terms of Type I error and one of the alternative methods may perform the best depending on the magnitude of θ_{GE}. For large-scale association studies, such as a genome-wide scan, it is expected that the gene-environment odds-ratios over different genes or/and exposures will have a distribution that has a large mass at or near the likely independence assumption, but would also include a range of values for θ_{GE} that corresponds to substantial violation of the independence assumption. In this article, we evaluate average power of alternative tests for interaction under some distributions for the genotype-exposure odds-ratio parameters that are likely to hold in large-scale association studies. These unique studies will help us to judge the overall performance of alternative statistical tests for interactions in light of the variations of the gene-environment association that are likely to appear in practice.

The article is organized as follows. In the section Different Tests for Interaction in a 2 × 4 Table, we first describe the different estimators that we consider. In the section Simulation Settings, we describe the simulation methods to evaluate the Type I error and power for different procedures. The section Data Examples presents data from the Colorectal Cancer (MECC) study [Poynter et al., 2005] to illustrate the behavior of the alternative estimators under a range of gene-environment association scenarios. In the section Simulation Findings, we present Type I error and power for these different estimators under different sample sizes, sampling ratios and varying strength of *G-E* association. The last section contains discussion and concluding remarks.

MATERIALS AND METHODS

DIFFERENT TESTS FOR INTERACTION IN A 2 × 4 TABLE

We consider the simple setup of an unmatched case-control study with a binary genetic factor *G* and a binary

TABLE I. Data for a unmatched case-control study with a binary genetic factor and a binary environmental exposure

	$G=0$		$G=1$		
	$E=0$	$E=1$	$E=0$	$E=1$	Total
$D=0$	r_{000}	r_{001}	r_{010}	r_{011}	n_0
$D=1$	r_{100}	r_{101}	r_{110}	r_{111}	n_1

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 genotype. Let D denote 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 Table I.

Let $\boldsymbol{r}_0=(r_{000}, r_{001}, r_{010}, r_{011})$ and $\boldsymbol{r}_1=(r_{100}, r_{101}, r_{110}, r_{111})$ denote the vector of observed cell frequencies in the controls and cases, respectively. The population parameters, namely, the cell probabilities corresponding to a particular G-E configuration in the underlying control and case populations are denoted as $\boldsymbol{p}_0=(p_{000}, p_{001}, p_{010}, p_{011}=1-p_{000}-p_{001}-p_{010})$ and $\boldsymbol{p}_1=(p_{100}, p_{101}, p_{110}, p_{111}=1-p_{100}-p_{101}-p_{110})$, respectively. The observed vectors of cell counts can be viewed as realizations from two independent multinomial distributions, namely, $\boldsymbol{r}_0 \sim \text{Multinomial}(n_0, \boldsymbol{p}_0)$ and $\boldsymbol{r}_1 \sim \text{Multinomial}(n_1, \boldsymbol{p}_1)$. Let $\text{OR}_{10}=p_{000}p_{101}/\ p_{001}p_{100}$ denote the odds-ratio associated with E for non-susceptible subjects ($G=0$), $\text{OR}_{01}=p_{000}p_{110}/\ p_{010}p_{100}$ denote the odds-ratio associated with G for unexposed subjects ($E=0$) and $\text{OR}_{11}=p_{000}p_{111}/\ p_{011}p_{100}$ denote the odds-ratio associated with $G=1$ and $E=1$ compared to the baseline category $G=0$ and $E=0$. Therefore,

$$\begin{aligned}\psi &= \text{OR}_{11}/(\text{OR}_{10}\text{OR}_{01}) \\ &= (p_{001}p_{010}p_{100}p_{111})/(p_{000}p_{011}p_{101}p_{110})\end{aligned}$$

is the multiplicative interaction parameter of interest.

In the following we describe the four estimators of interaction considered in the article.

1. *The case-control estimator*: The classical estimator of the interaction log-odds-ratio, namely, $\log(\psi)=\beta$, obtained from case-control data is given by

$$\hat{\beta}_{\text{CC}} = \log\left(\frac{r_{001}r_{010}r_{100}r_{111}}{r_{000}r_{011}r_{101}r_{110}}\right).$$

Note that $\hat{\beta}_{\text{CC}}$ is the maximum likelihood estimate (MLE) of β based on the likelihood of the data given in Table I allowing for *any* valid joint distribution for G and E, without any constraints like gene-environment independence. Employing standard asymptotic theory, variance of the case-control estimator can be estimated as $\hat{\sigma}^2_{\text{CC}}=\sum_{d=0}^{1}\sum_{g=0}^{1}\sum_{e=0}^{1}(1/r_{dge})$. We will use the Wald test for interaction based on the standardized Z statistic $Z_{\text{CC}}=\hat{\beta}_{\text{CC}}/\hat{\sigma}_{\text{CC}}$.

2. *The case-only estimator*: The odds-ratio interaction parameter ψ can be expressed as a ratio of two odds-ratios, namely,

$$\psi = \frac{\text{Odds-ratio between } G \text{ and } E \text{ among cases}}{\underbrace{\text{Odds-ratio between } G \text{ and } E \text{ among controls}}_{=1 \text{ under } G-E \text{ independence and rare disease}}}. \tag{1}$$

The denominator in ψ, namely, the population odds-ratio between G and E among the disease-free subjects, reduces to unity under gene-environment independence and rare disease assumption. Thus, under those two assumptions, ψ can be unbiasedly estimated by the sample odds-ratio between G and E among the cases alone [Piergorsch et al., 1994]. The case-only estimator gains efficiency over its case-control counterpart by reduction of the variance associated with estimation of the odds-ratio between G and E among the controls.

More formally, the case-only estimator of the interaction log-odds-ratio is given by

$$\hat{\beta}_{\text{CO}} = \log\left(\frac{r_{100}r_{111}}{r_{101}r_{110}}\right).$$

It has been shown that $\hat{\beta}_{\text{CO}}$ is the MLE of β under the constraint of G-E independence in control population [Umbach and Weinberg, 1997]. The asymptotic variance of the case-only estimator can be obtained as $\hat{\sigma}^2_{\text{CO}}=\sum_{g=0}^{1}\sum_{e=0}^{1}(1/r_{1ge})$. The test for interaction associated with the case-only procedure is again based on the asymptotic normality of the constrained MLE, using the standardized Z statistic $Z_{\text{CO}}=\hat{\beta}_{\text{CO}}/\hat{\sigma}_{\text{CO}}$.

The case-only estimator is unbiased under G-E independence assumption and has a much reduced variance compared to the case-control estimator; note that $\hat{\sigma}^2_{\text{CO}}<\hat{\sigma}^2_{\text{CC}}$. As a result, the tests based on the case-only estimator have significantly enhanced power to detect gene-environment interaction, compared to their case-control counterparts. However, the case-only estimator is subject to potential bias under departures from gene-environment independence assumption. From the representation in (1), for example, it is clear that if gene-environment independence does not hold, i.e. when the odds-ratio in the denominator of (1) departs from unity, the case-only estimator of the interaction parameter will remain asymptotically biased by a magnitude that is exactly equal to the G-E odds-ratio in the control population. This leads to a highly inflated Type I-error rate for the corresponding case-only testing procedure.

3. *The two-step estimator*: A measure of G-E association in the control population is given by the log-odds-ratio between G and E among subjects with $D=0$, namely,

$$\theta_{GE} = \log\{(p_{000}p_{011})/(p_{001}p_{010})\}. \tag{2}$$

The assumption of G-E independence, together with the rare disease approximation, implies, $\theta_{GE}=0$. The MLE of θ_{GE} is given by $\hat{\theta}_{GE}=\log\{(r_{000}r_{011})/(r_{001}r_{010})\}$ with an estimate of the asymptotic variance given by $\hat{\sigma}^2_{\theta_{GE}}=\sum_{g=0}^{1}\sum_{e=0}^{1}(1/r_{0ge})$. One could use $\hat{\theta}_{GE}$ to first test the hypothesis $H_0:\theta_{GE}=0$. If the null hypothesis is rejected, one could then use the case-control estimator, and if one fails to reject the null hypothesis of gene-environment independence, one could proceed to use the case-only estimator.

More formally, the two-step procedure [Albert et al., 2001] tests for G-E independence in the control population, namely, $H_0:\theta_{GE}=0$, at a chosen level of significance α

using the test statistic $Z_{GE} = \hat{\theta}_{GE}/\hat{\sigma}_{\theta_{GE}}$. If $|Z_{GE}| > Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the upper $100(1\alpha/2)$-th percentile of the standard normal distribution, one uses Z_{CC} and if $|Z_{GE}| \leq Z_{\alpha/2}$ the test is based on Z_{CO}. The two-step test statistic can be expressed in a concise form as

$$Z_{\text{TS}} = Z_{\text{CC}} I[|Z_{GE}| > Z_{\alpha/2}] + Z_{\text{CO}} I[|Z_{GE}| \leq Z_{\alpha/2}],$$

where $I[A]$ is the indicator function if A holds and is zero otherwise.

4. *EB type shrinkage estimator*: We have recently proposed a new estimator of the interaction parameter, which attempts to relax the *G-E* independence assumption in a data-adaptive way (MC). The proposed method involves a weighted combination of the case-control and the case-only estimator in the form of

$$\hat{\beta}_{\text{EB}} = \frac{\hat{\sigma}^2_{\text{CC}}}{(\hat{\theta}^2_{GE} + \hat{\sigma}^2_{\text{CC}})} \hat{\beta}_{\text{CO}} + \frac{\hat{\theta}^2_{GE}}{(\hat{\theta}^2_{GE} + \hat{\sigma}^2_{\text{CC}})} \hat{\beta}_{\text{CC}}. \qquad (3)$$

The EB perspective for constructing this estimator is described in detail in MC and is motivated by Greenland [1993]. We note that although it was constructed from a Bayesian perspective, this estimator is purely a functional of data depending on four ingredients: $\hat{\beta}_{\text{CO}}$, $\hat{\beta}_{\text{CC}}$, $\hat{\sigma}_{\text{CC}}$ and $\hat{\theta}_{GE}$, which are all functions of the cell counts in the 2×4 table.

To understand the intuitive rationale behind the estimator, observe that as $\hat{\theta}_{GE} \to 0$, i.e. as the data provide evidence in favor of *G-E* independence, $\hat{\beta}_{\text{EB}} \to \hat{\beta}_{\text{CO}}$, and as $\hat{\theta}_{GE} \to \infty$, i.e. as the uncertainty regarding *G-E* independence in control population becomes stronger, $\hat{\beta}_{\text{EB}} \to \hat{\beta}_{\text{CC}}$. Also, when the true $\theta_{GE} \neq 0$, i.e. the independence assumption is violated, then as the sample size $n \to \infty$, $\sigma^2_{\text{CC}} \to 0$ and hence $\hat{\beta}_{\text{EB}} \to \hat{\beta}_{\text{CC}}$, the unbiased case-control estimator.

MC used Taylor's approximation to propose an estimator of the variance of $\hat{\beta}_{\text{EB}}$ in the form

$$\widehat{V}_A(\hat{\beta}_{\text{EB}}) \approx \hat{\sigma}^2_{\text{CO}} + \left(\frac{\hat{\theta}^2_{GE}(\hat{\theta}^2_{GE} + 3\hat{\sigma}^2_{\text{CC}})}{(\hat{\sigma}^2_{\text{CC}} + \hat{\theta}^2_{GE})^2} \right)^2 \hat{\sigma}^2_{\theta_{GE}}. \qquad (4)$$

In the simulation results presented by MC, this variance approximation was shown to work fairly well even for smaller sample sizes. Figure 1 in the supplementary material presents the sampling distribution of the EB estimator, which appears to be fairly normal though the asymptotic distribution theory for this shrinkage estimator needs to be rigorously established. We will base the EB test for the interaction based on the Wald statistic $Z_{\text{EB}} = \hat{\beta}_{\text{EB}}/\sqrt{\widehat{V}_A(\hat{\beta}_{\text{EB}})}$.

In this article, we also consider a slightly modified version of the EB estimator proposed by MC. The interaction log-odds-ratio β in a 2×4 table is given by

$$\beta = \beta_{\text{CO}} - \theta_{GE}. \qquad (5)$$

The case-only estimator is obtained from (5) by assuming $\theta_{GE} = 0$, whereas the case-control estimator is obtained by substituting $\theta_{GE} = \hat{\theta}_{GE}$ (with β_{CO} estimated by its sample counterpart $\hat{\beta}_{\text{CO}}$). Another possible estimation strategy with Bayesian flavor is to estimate θ_{GE} by a weighted estimator of the prior guess 0, reflecting gene-environment independence and the empirical estimate of gene-environment association, namely, $\hat{\theta}_{GE}$. An EB estimate of θ_{GE}, following arguments similar to those for the construction of β_{EB} described in MC, is given by $\hat{\theta}_{\text{EB}} = \{1 + (\hat{\sigma}^2_{\theta_{GE}}/\hat{\theta}^2_{GE})\}^{-1}\hat{\theta}_{GE}$. Substituting θ_{GE} by $\hat{\theta}_{\text{EB}}$ in (5), we obtain an estimate of the interaction parameter, which can be also expressed as a weighted average of the case-only and case-control estimators with weights of similar form as those in (3), except that is $\hat{\sigma}^2_{\text{CC}}$ is replaced with $\hat{\sigma}^2_{\theta_{GE}}$. We will label this modified estimator as $\hat{\beta}_{\text{EB2}}$. The variance expression for $\hat{\beta}_{\text{EB2}}$, $\hat{V}_A(\hat{\beta}_{\text{EB2}})$, say, is exactly of the same form as in (4) but replacing $\hat{\sigma}^2_{\text{CC}}$ by $\hat{\sigma}^2_{\theta_{GE}}$. A Wald-type test is again constructed based on the statistic $Z_{\text{EB2}} = \hat{\beta}_{\text{EB2}}/\sqrt{\hat{V}_A(\hat{\beta}_{\text{EB2}})}$.

SIMULATION SETTINGS

In our simulation, we first investigate the Type I error and power of these four different testing procedures under various alternative values of β across a spectrum of association scenarios for G and E and varying sample sizes and sampling ratios. All Type I error and power calculations are based on simulated data sets under different parameter settings.

In the fixed parameter setting, we fix the values for the prevalences of G and E, namely, P_G and P_E, and the value of the odds-ratio θ_{GE} in the control population. Fixing these three quantities, one is able to obtain the control probability vector $\boldsymbol{p}_0$ by solving the following system of equations:

$$\theta_{GE} = \frac{p_{000}(p_{000} - (1 - P_G - P_E))}{(1 - P_G - p_{000})(1 - P_E - p_{000})},$$

$$p_{001} = 1 - P_G - p_{000},$$

$$p_{010} = 1 - P_E - p_{000}.$$

We then set the values of OR_{10}, OR_{01} and ψ, which together with $\boldsymbol{p}_0$ define the case-probability vector [Satten and Kupper, 1993]. We generate data independently from the two multinomial distributions corresponding to the case and control populations. We then compute the case-control, case-only, two-step and the two proposed EB-type estimators, their standard errors and the corresponding Z statistics. The Z statistics are then compared with the critical value from the standard normal distribution for a given α. Type I error and power are then estimated by the proportion of null hypotheses rejected at a given level of significance, i.e. the proportion of times $|Z| > Z_{\alpha/2}$ for each method in 10,000 replications. We considered prevalence values of G and E, namely, $P_G = P_E = 0.3$, and values of θ_{GE} in the range of 0–log(2). For the disease-risk parameters, we consider a setting with no main effects ($\text{OR}_{10} = \text{OR}_{01} = 1$) and varying values of β again in the range of $0-\log(2)$. We consider two levels of significance 5 and 0.5% and number of cases $n_1 = 500$ and 1,000.

We would now like to emphasize two major aspects of our simulation study.

1. *Effect of varying sampling ratio*: In this article we assess the effect of varying number of controls while fixing the number of cases on alternative tests for gene-environment interaction. In the case-only analysis, controls do not play any role in the inference on the interaction parameter. In a traditional case-control study, one typically tries to maintain approximately the same number of controls as the

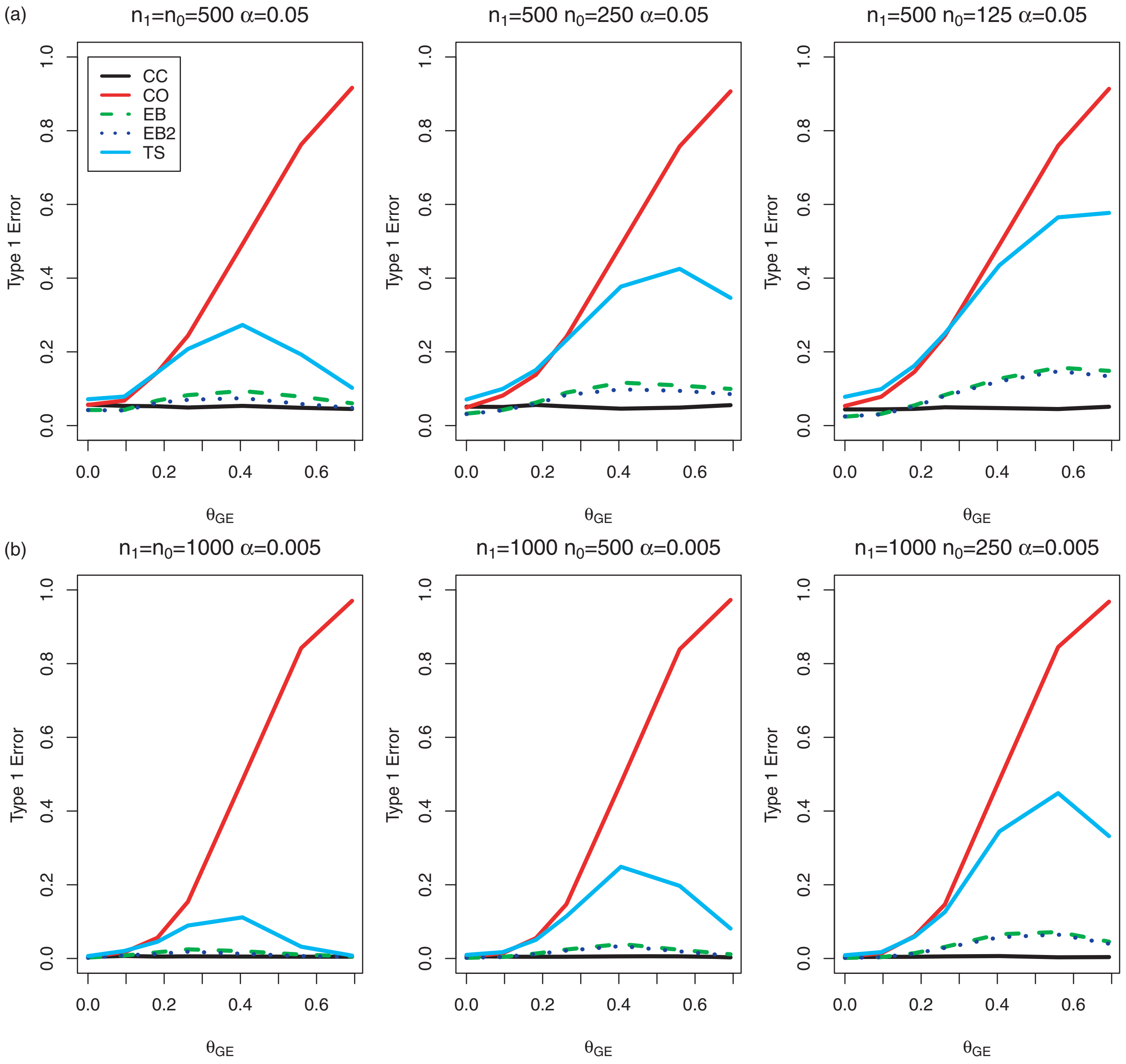

Fig. 1. Type I-error rates for different estimators under two settings: (a) 500 cases with varying number of controls for $\alpha = 0.05$ and (b) 1,000 cases with varying number of controls for $\alpha = 0.005$. The horizontal axis in each plot represents the true *G-E* log-odds-ratio among controls, namely, θ_{GE}. We consider $P_G = P_E = 0.3$, $\mathrm{OR}_{10} = \mathrm{OR}_{01} = 1$ in all settings. Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.

number of cases. In the current simulation, we consider intermediate study designs where we vary the case:control sampling ratio from 1:1, 2:1, 4:1. From a design perspective, we investigate (a) the loss of power in alternative methods due to recruitment of fewer number of controls and (b) the ability of the methods to maintain a desired Type I-error level with decreasing number of controls.

2. *Evaluating integrated Type I error and power*: We assess the Type I error and power of alternative procedures averaged over a distribution of θ_{GE} that might reflect varying scenarios of gene-environment association in large-scale studies. In dealing with thousands of genes and their interactions with environmental factors, one is more interested in assessing the *average* performance of the methods across the whole ensemble of gene-environment configurations, rather than a specific single parameter setting. In particular, we consider a mixture distribution of θ_{GE} that assigns 80% weight to a point mass at zero and 20% weight on a normal distribution centered at zero and an SD of $\log(1.5)/2$. The scenario, for example, may correspond to a study where a large variety of genotype-exposure combinations are being studied, and for 80% of those combinations the gene-environment independence assumption is satisfied and for the rest the genotype-exposure log-odds-ratios have a distribution with 95% mass within $\pm\log(1.5)$ limits. To enumerate average

performance of the different procedures, we first enumerate their Type I error and power on a fixed grid of values for θ_{GE} and then take weighted average of those values with weights obtained from the specified mixture distribution.

RESULTS

DATA EXAMPLES

The MECC study is a population-based case-control study of patients who received a diagnosis of invasive colorectal cancer in northern Israel between March 31, 1998, and March 31, 2004. Controls were matched according to age, sex, clinic and ethnic group (Jewish vs. non-Jewish). Participants were interviewed to obtain demographic information, personal and family history of cancer, medical history, medication use and health habits. They also completed a dietary questionnaire and a blood sample was collected. Dietary data were collected from in-person interviews using a 187-item semi-quantitative food-frequency questionnaire (FFQ). The FFQ was based on the instrument originally developed by Willett and Reynolds [1987], and was expanded and tailored to the Israeli diet. The semi-quantitative FFQ was assessed for reliability and validity with repeated measures and validation against a 3-day dietary diary with good correspondence.

Genomic DNA was extracted from blood using the Puregene kit (Gentra Systems, Inc., Minneapolis, MN). Genotyping was performed for glutathione *S*-transferase M1 (GSTM1) and *N*-acetyl transferase type 2 (NAT2) by allele-specific polymerase chain reaction and polymerase chain reaction-restriction fragment length polymorephism, respectively. GSTM1 genotyping results permitted dichotomous classification as GSTM1 null or non-null. NAT2 genotyping was used to classify individuals as fast or slow acetylators for consistency with previously published literature [Roberts-Thomson et al., 1996]. We study the interaction between GSTM1 and NAT2 with smoking and several measures of dietary consumption to illustrate our different methods. Analysis in each case was based on complete case data for each gene-diet configuration. The observed cell counts for each 2×4 configuration is presented in Table II. The dichotomization scheme for the genetic and behavioral exposures is described at the bottom of Table II.

The acetylator phenotype has been hypothesized to modulate the relationship between red meat and risk of colorectal cancer (CRC), with increasing red meat consumption associated with increased risk of CRC among fast, but not slow, acetylators in some [Roberts-Thomson et al., 1996], but not all, studies [Barrett et al., 2003]. NAT2 has also been suggested as a potential modifier of the relationship between cigarette smoking and CRC as well as colorectal adenomas, although studies are not consistent [Moslehi et al., 2006; Barrett et al., 2003].

Homozygous deletion of the GSTM1 allele corresponds to a null phenotype that has been suggested to modify the well-known protective association between increasing cruciferous vegetable consumption and decreased risk of CRC [Lin et al., 1998]. However, this relationship is also inconsistent in the literature. Since GSTM1 also metabolizes components of tobacco smoke, GSTM1 has been studied as a modifier of the relationship between smoking and CRC. However, meta-analyses do not support an interaction [Smits et al., 2003].

TABLE II. The 2×4 tables for studying several gene-environment interactions in the Molecular Epidemiology of Colorectal Cancer (MECC) study[a]

		$G = 0$		$G = 1$		
		$E = 0$	$E = 1$	$E = 0$	$E = 1$	Total
NAT2∗Smoking	$D = 0$	623	540	398	266	1,827
	$D = 1$	607	479	397	280	1,763
NAT2∗Beef	$D = 0$	867	291	487	176	1,821
	$D = 1$	798	234	474	176	1,682
GSTM1∗Smoking	$D = 0$	808	118	765	124	1,815
	$D = 1$	746	58	796	81	1,681
GSTM1∗Beef	$D = 0$	735	275	781	240	2,031
	$D = 1$	727	225	650	223	1,825
GSTM1∗Vegetable	$D = 0$	210	801	238	788	2,037
	$D = 1$	258	705	266	620	1,849
GSTM1∗Fruit	$D = 0$	228	783	232	792	2,035
	$D = 1$	244	720	258	632	1,854

[a] For each configuration of *G* and *E*, observations with missing data on either *G* and *E* were deleted.
Dichotomization schemes for *G* and *E*:
NAT2 (fast vs. slow) and Smoking (ever vs. never).
NAT2 (fast vs. slow) and Beef (upper 25% vs. lower 75%).
GSTM1 (non-null vs. null) and Smoking (current vs. former).
GSTM1 (non-null vs. null) and Beef (upper 25% vs. lower 75%).
GSTM1 (non-null vs. null) and Vegetable (upper 75% vs. lower 25%).
GSTM1 (non-null vs. null) and Fruit (upper 25% vs. lower 75%).
NAT2, *N*-acetyl transferase type 2; GSTM1, glutathione *S*-transferase M1.

One can notice the adaptive feature of the EB-type estimators depending on the strength of *G-E* association across the six different cases. For example, while estimating GSTM1∗fruit interaction, $\hat{\theta}_{GE} = 0.00$ (*P*-value = 0.96), providing strong evidence in favor of *G-E* independence, the EB estimator of interaction log-odds-ratio $\hat{\beta}_{\mathrm{EB}} = -0.19$ (*P*-value = 0.08, CI = (−0.39,0.02)) is identical to the case-only estimator. The results suggest possible effect modification of fruit consumption on the risk of CRC by GSTM1 genotype status. However, the case-control analysis in this context does not detect this marginal interaction effect ($\hat{\beta}_{\mathrm{CC}} = -0.18$, *P*-value = 0.23, CI = (−0.47,0.11)).

In contrast, for estimation of GSTM1∗beef interaction where $\hat{\theta}_{GE} = -0.20$ (*P*-value = 0.05), presenting evidence against the independence assumption, EB inference regarding the interaction log-odds-ratio ($\hat{\beta}_{\mathrm{EB}} = 0.23$, *P*-value = 0.14, CI = (−0.08,0.53)) is closer to case-control inference ($\hat{\beta}_{\mathrm{CC}} = 0.30$, *P*-value = 0.05, CI = (−0.01,0.59)) than the case-only inference ($\hat{\beta}_{\mathrm{CO}} = 0.10$, *P*-value = 0.34, CI = (−0.11,0.32)). Similarly, under a strong violation of the independence assumption between NAT2 and smoking ($\hat{\theta}_{GE} = -0.26$, *P*-value = 0.01), EB estimate regarding the interaction parameter ($\hat{\beta}_{\mathrm{EB}} = 0.09$, *P*-value = 0.55, CI = (−0.20, 0.38)) resembles the case-control estimate ($\hat{\beta}_{\mathrm{CC}} = 0.15$, *P*-value = 0.29, CI = (−0.13,0.42)), while the case-only estimate is in the entirely opposite direction ($\hat{\beta}_{\mathrm{CO}} = -0.11$, *P*-value = 0.26, CI = (−0.31,0.08)).

TABLE III. The results from the MECC study on analyzing interaction effects of NAT2 and GSTM1 with smoking and dietary exposures on the risk of colorectal cancer

		Interaction log-odds-ratio β				
		Estimate	SE	*P*-value	LCI	UCI
NAT2∗Smoking	Case-only	−0.11	0.10	0.26	−0.31	0.08
	Case-control	0.15	0.14	0.29	−0.13	0.42
$\hat{\theta}_{GE} = -0.26$	Two-step	0.15	0.14	0.29	−0.13	0.42
P-value = 0.01	EB	0.09	0.15	0.55	−0.20	0.38
	EB2	0.11	0.15	0.43	−0.17	0.40
NAT2∗Beef	Case-only	0.24	0.12	0.04	0.01	0.46
	Case-control	0.16	0.16	0.31	−0.15	0.48
$\hat{\theta}_{GE} = 0.07$	Two-step	0.24	0.12	0.04	0.01	0.46
P-value = 0.51	EB	0.22	0.13	0.08	−0.02	0.47
	EB2	0.21	0.14	0.13	−0.06	0.49
GSTM1∗Smoking	Case-only	0.27	0.18	0.13	−0.08	0.62
	Case-control	0.16	0.23	0.47	−0.28	0.61
$\hat{\theta}_{GE} = 0.10$	Two-step	0.27	0.18	0.13	−0.08	0.62
P-value = 0.45	EB	0.25	0.19	0.19	−0.123	0.62
	EB2	0.23	0.21	0.28	−0.19	0.65
GSTM1∗Beef	Case-only	0.10	0.11	0.34	−0.11	0.32
	Case-control	0.30	0.15	0.05	0.01	0.59
$\hat{\theta}_{GE} = -0.20$	Two-step	0.10	0.11	0.34	−0.11	0.32
P-value = 0.05	EB	0.23	0.16	0.14	−0.08	0.53
	EB2	0.26	0.16	0.10	−0.05	0.57
GSTM1∗Vegetable	Case-only	−0.16	0.10	0.12	−0.36	0.04
	Case-control	−0.02	0.15	0.91	−0.31	0.27
$\hat{\theta}_{GE} = -0.14$	Two-step	−0.16	0.10	0.12	−0.36	0.04
P-value = 0.19	EB	−0.09	0.15	0.53	−0.38	0.20
	EB2	−0.07	0.16	0.66	−0.38	0.24
GSTM1∗Fruit	Case-only	−0.19	0.10	0.08	−0.39	0.02
	Case-control	−0.18	0.15	0.23	−0.47	0.11
$\hat{\theta}_{GE} = -0.01$	Two-step	−0.19	0.10	0.08	−0.39	0.02
P-value = 0.96	EB	−0.19	0.10	0.08	−0.39	0.02
	EB2	−0.19	0.10	0.08	−0.39	0.02

The sample control odds-ratio between G and E, namely, $\hat{\theta}_{GE}$ and the P-value for testing $H_0 : \theta_{GE} = 0$ are presented in the first column. Also included are the point estimate, corresponding large sample standard error, P-value for testing $H_0 : \beta = 0$ and the 95% CI for the interaction log-odds-ratio parameter β.

MECC, Molecular Epidemiology of Colorectal Cancer; NAT2, *N*-acetyl transferase type 2; GSTM1, glutathione *S*-transferase M1; EB, empirical Bayes.

The other three scenarios in Table III, corresponding to NAT2∗beef interaction, GSTM1∗smoking interaction and GSTM1∗vegetable interaction, all represent data scenarios where there is not much evidence against the gene-environment independence assumption. There is very little evidence against independence, for example, between NAT2 and beef consumption (($\hat{\theta}_{GE} = 0.07$, P-value = 0.51). In this case, EB inference regarding interaction ($\hat{\beta}_{\text{EB}} = 0.22$, P-value = 0.08, CI = (−0.03, 0.47)) is quite close to case-only inference ($\hat{\beta}_{\text{CO}} = 0.24$, P-value = 0.04, CI = (0.01,0.46)), both suggesting modest evidence of non-multiplicative interaction. The case-control estimator cannot detect any evidence of such an interaction.

The various association and interaction scenarios considered in the above real data example serve as an illustration of how inference could drastically change by assuming gene-environment independence when it is not true and how the conclusions from case-control and case-only analysis could be widely different depending on the *G-E* association present in a particular data situation. The examples clearly emphasize the need and usefulness of a data-adaptive compromise where one does not have to specify or rely on unverifiable model assumptions and simply let the data determine the evidence in favor of or against the independence assumption. However, as we will note in the following section, one is still able to maintain significant efficiency advantages without losing much of the desired robustness properties.

SIMULATION FINDINGS

We now summarize the main findings in the simulation setting mentioned above. Table IV and Figure 1 present the Type I-error values for testing $H_0 : \beta = 0$ at $n_1 = 500$, $\alpha = 0.05$ and $n_0 = 1,000$, $\alpha = 0.005$, respectively, for all the five methods under a two-sided alternative and varying values of θ_{GE}. One can notice the highly inflated Type I-error levels for the case-only method and also the two-step procedure under violation of *G-E* independence. The Type I-error inflation of the case-only method can be noticed even under very small departures from the independence assumption, say when $\theta_{GE} = \log(1.1)$, $\alpha = 0.05$ and $n_1 = 500$, the Type I error of CO is 0.077. With $\theta_{GE} = \log(1.2)$, $\alpha = 0.05$ and $n_1 = 500$, the Type I error for CO is 0.143, much above the nominal level. With increase in θ_{GE} to log(1.5), the Type I error for CO reaches 0.498, which is unacceptable for any testing procedure. The two-step procedure often also has unacceptably high Type I error. However, the Type I error for the TS procedure, unlike that of CO, does decrease with larger values of θ_{GE} and increasing sample size. For example, when $n_1 = 500$, $\alpha = 0.05$ and sampling ratio of 1:1, Type I error for TS is 0.278 when $\theta_{GE} = \log(1.5)$, but reduces to 0.111 when $\theta_{GE} = \log(2.0)$. The effect of sample size on Type I-error rates for different methods has been reported in Table I of supplementary materials. Increasing sample size does not reduce the Type I error of case-only procedure; however, the Type I error of TS procedure does decrease with increasing sample size (compare the results of Table IV in the text with $n_1 = 500$, $\alpha = 0.05$ to the results of Table I in the supplementary materials with $n_1 = 1,000$, $\alpha = 0.05$).

The EB-type estimators provide a much better control of Type I error, especially under smaller departures from the independence assumption. For example, when $n_1 = n_0 = 500$ and $\alpha = 0.05$, the EB procedure maintains close to nominal level of Type I error (0.052) at $\theta_{GE} = \log(1.1)$. The Type I error for EB increases to 0.076 at $\theta_{GE} = \log(1.2)$, and then to 0.097 for $\theta_{GE} = \log(1.5)$ and eventually drops to 0.059 when $\theta_{GE} = \log(2.0)$. Increasing sample size reduces Type I error for EB procedures (see Table I in supplementary material). The EB2 method has consistently slightly smaller Type I error than EB. Figure 2 shows the basic variation pattern of Type I error with changes in θ_{GE} for all five estimators considered. Note that while the TS procedure has similar behavior patterns like EB procedure, the Type I curve lies much above EB for all simulation settings.

TABLE IV. Type I error for different estimators under two settings: (a) 500 cases with varying number of controls when $\alpha = 0.05$ and (b) 1,000 cases with varying number of controls when $\alpha = 0.005$

		$\alpha = 0.05$, $n_1 = 500$					$\alpha = 0.005$, $n_1 = 1,000$				
θ_{GE}	$n_1 : n_0$	CC	CO	EB	EB2	TS	CC	CO	EB	EB2	TS
0	1:1	0.052	0.051	0.037	0.035	0.066	0.006	0.004	0.004	0.005	0.009
	2:1	0.053	0.051	0.032	0.033	0.073	0.005	0.004	0.003	0.003	0.007
	4:1	0.048	0.051	0.023	0.023	0.075	0.003	0.004	0.003	0.002	0.009
log(1.1)	1:1	0.052	0.077	0.052	0.044	0.091	0.005	0.016	0.006	0.005	0.018
	2:1	0.045	0.077	0.040	0.036	0.094	0.005	0.016	0.004	0.004	0.019
	4:1	0.051	0.077	0.029	0.029	0.099	0.006	0.016	0.003	0.003	0.019
log(1.2)	1:1	0.057	0.143	0.076	0.070	0.151	0.004	0.060	0.016	0.013	0.047
	2:1	0.048	0.143	0.064	0.060	0.159	0.004	0.060	0.013	0.011	0.050
	4:1	0.050	0.143	0.049	0.050	0.155	0.005	0.060	0.015	0.014	0.057
log(1.5)	1:1	0.044	0.498	0.097	0.075	0.278	0.006	0.472	0.019	0.014	0.110
	2:1	0.050	0.498	0.113	0.098	0.381	0.006	0.472	0.046	0.038	0.242
	4:1	0.051	0.498	0.139	0.132	0.447	0.005	0.472	0.066	0.060	0.334
log(2.0)	1:1	0.047	0.914	0.059	0.049	0.111	0.005	0.973	0.006	0.005	0.008
	2:1	0.049	0.914	0.092	0.078	0.342	0.006	0.973	0.014	0.010	0.084
	4:1	0.048	0.914	0.157	0.145	0.582	0.005	0.973	0.046	0.039	0.346

We consider $P_G = P_E = 0.3$, $OR_{10} = OR_{01} = 1$ in all settings.
CC, case-control; CO, case-only; EB, empirical Bayes; TS, two-step.

Figure 1 and Table IV also illustrate the effect of sampling ratios on Type I-error rates of the different procedures. One can notice that even with recruiting a smaller number of controls (half or one-fourth the number of cases) and using the adaptive EB-type estimators, one can prevent the Type I-error inflation of the CO method to a large extent. For example, for $n_1 = 1,000$, $\alpha = 0.005$, $\theta_{GE} = \log(1.1)$, the Type I error for CO is 0.016, whereas the Type I error for EB with sampling ratio 2:1 is 0.004. The TS procedure does much worse in achieving the same goal with the same number of controls when compared to the EB methods (TS Type I error = 0.019 under same situation).

Table V presents some representative numerical values for power with interaction odds-ratio fixed at 1.5 under settings identical to the ones considered in Table IV. Tables IV and V should be evaluated simultaneously to reiterate that the extremely impressive gain in power for the case-only method comes at a cost of Type I-error values much higher than the nominal level. Under $\theta_{GE} = 0$, $\alpha = 0.005$, $n_1 = 1,000$, CO is surely the preferred choice with power 0.52 compared to 0.19 for CC, both maintaining nominal level of Type I error. In the same setting, EB maintains significant gain in terms of power when compared to CC with a power of 0.35. EB2 always has slightly less power than EB. Note that, under modest violation of *G-E* independence assumption, with $\theta_{GE} = \log(1.1)$, $\alpha = 0.005$, $n_1 = 1,000$, the power of CO is 0.77, but accompanied with a Type I error of 0.016, which is 3 times more than the designated level of significance of 0.005. In the same setting with a sampling ratio of 1:1, the power of EB is 0.46, but Type I error is very close to the nominal level (0.006). The CC method has power 0.20 (Type I error 0.004) in the same setting. Under modest departure of independence assumption ($\theta_{GE} \leq \log(1.2)$), EB continues to maintain this power gain over CC. However, for larger values of θ_{GE}, EB puts most of its weight on CC and thus the power of EB becomes closer to the power of CC.

We next study the effect of sampling ratios on the power. It is interesting to compare the power of the EB method to the case-control analysis when the sampling ratio for controls decreases. Under $\theta_{GE} = \log(1.1)$, for $\alpha = 0.05$, $n_1 = 500$ and a sampling ratio of 2:1, the power for EB is 0.43 (Type I error 0.04), which is more than twice the power of CC (power = 0.20, Type I error = 0.05) with the same sample size and is even higher than the power obtained by CC with double the number of controls (power for CC = 0.30, Type I error = 0.05, sampling ratio 1:1). This suggests that one can detect given effect sizes with higher level of power by recruiting fewer controls when compared to standard case-control analysis, just by adopting the EB testing procedure. Also note that the rate of decrease in power with a decrease in control sampling rates is slower for the EB procedure than the corresponding case-control analysis. In Table V, for example, when $n_1 = 500$, $\alpha = 0.05$ and $\beta = \log(1.5)$, as the sampling ratio changes from 1:1 to 2:1 to 4:1 the power for EB changes from 0.50 to 0.43 and then to 0.32, reflecting successive percentage reduction in power of 14 and 25.6%, respectively. The corresponding power values for CC decrease from 0.30 to 0.20 and then to 0.13 with the percentage reductions in power being 50 and 35%, respectively. Figures 2 and 3 present power curves corresponding to the different testing approaches under the two settings: (a) $\alpha = 0.05$ and $n_1 = 500$ and (b) $\alpha = 0.005$ and $n_1 = 1,000$ with the number of controls varying across each row of the graphical array and with values of θ_{GE} varying across the columns of the array. One can notice the marked power gains of the EB estimator relative to the case-control estimator under modest departures from the gene-environment independence assumption as well as the power advantage when control:case ratio falls below 1:1.

We now draw our attention to the results on integrated Type I error and power in Table VI, integrated with respect to a mixture distribution as discussed in the section simulation settings.

Under the mixture distribution setting considered in Table VI, both EB methods maintain Type I error very well for all scenarios. For example, with $n_1 = n_0 = 1,000$,

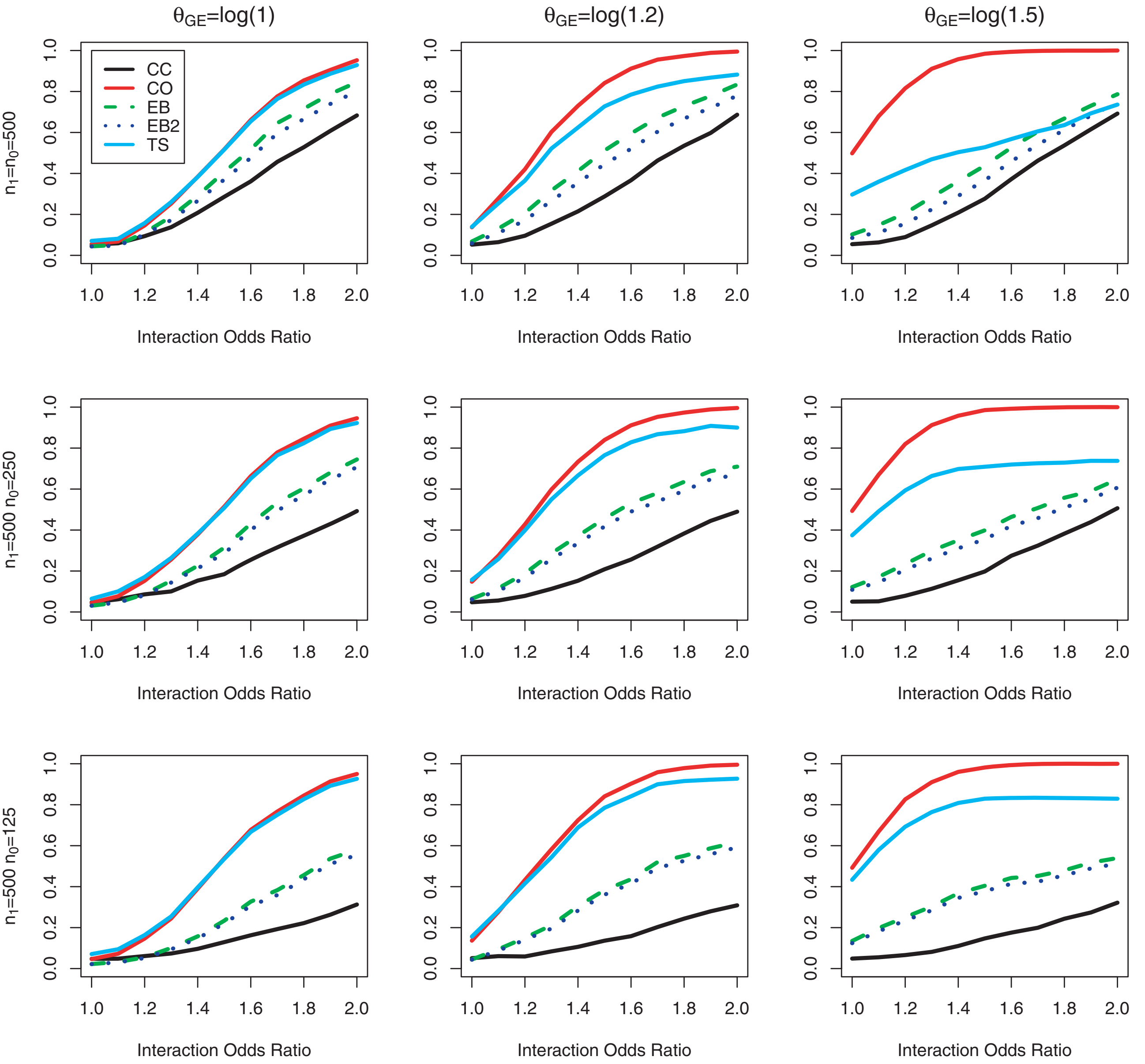

Fig. 2. Power for different estimators with 500 cases and number of controls varying across each row of the graphical array. The true value for θ_{GE} is set at 0, log(1.2) and log(1.5) across each column of the array. The horizontal axis in each plot represents the true value of the interaction odds ratio $\psi = \exp(\beta)$. The Type 1 error level is set at $\alpha = 0.05$. We consider $P_G = P_E = 0.3$, $\mathrm{OR}_{10} = \mathrm{OR}_{01} = 1$ in all settings. Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.

$\alpha = 0.005$ and $\beta = 0$, the Type I error for EB is 0.004, while CO has a Type I error of 0.021. The power of the EB estimators is generally much higher than that of CC with sample sizes being equal. In fact, the power for EB is comparable or higher than CC even when the latter method uses twice the number of controls. For example, when $n_1 = n_0 = 1,000$, $\alpha = 0.005$ and $\beta = \log(2.0)$, the power of CC is 0.726. With half the number of controls, i.e. $n_0 = 500$, the power of EB is 0.741. As noted in the fixed parameter setting of Table V, the rate of decrease in power with the decreasing sampling rate for controls is slower for EB than that for CC.

Results for several other simulation settings are available in the supplementary material accompanying the article.

DISCUSSION

In summary, our study indicates that the novel EB-type shrinkage estimation procedure leads to a promising method for testing gene-environment interaction in case-control studies. The method can gain major power over standard case-control analysis by exploiting the likely

TABLE V. Power at $\beta = \log(1.5)$ for different estimators under two settings: (a) 500 cases with varying number of controls when $\alpha = 0.05$ and (b) 1,000 cases with varying number of controls when $\alpha = 0.005$

		$\alpha = 0.05$, $n_1 = 500$					$\alpha = 0.005$, $n_1 = 1,000$				
θ_{GE}	$n_1 : n_0$	CC	CO	EB	EB2	TS	CC	CO	EB	EB2	TS
0	1:1	0.29	0.53	0.41	0.39	0.52	0.19	0.52	0.35	0.31	0.52
	2:1	0.20	0.53	0.33	0.30	0.52	0.11	0.52	0.26	0.23	0.51
	4:1	0.14	0.53	0.23	0.21	0.53	0.05	0.52	0.18	0.17	0.53
log(1.1)	1:1	0.30	0.71	0.50	0.45	0.66	0.20	0.77	0.46	0.39	0.70
	2:1	0.20	0.71	0.43	0.39	0.68	0.11	0.77	0.38	0.33	0.73
	4:1	0.13	0.71	0.32	0.30	0.69	0.06	0.77	0.28	0.26	0.73
log(1.2)	1:1	0.29	0.84	0.51	0.45	0.72	0.21	0.92	0.43	0.35	0.71
	2:1	0.21	0.84	0.45	0.41	0.77	0.11	0.92	0.38	0.34	0.79
	4:1	0.13	0.84	0.38	0.35	0.78	0.05	0.92	0.34	0.31	0.83
log(1.5)	1:1	0.29	0.98	0.45	0.38	0.54	0.21	1.00	0.31	0.25	0.31
	2:1	0.20	0.98	0.40	0.35	0.70	0.11	1.00	0.24	0.21	0.49
	4:1	0.14	0.98	0.4.0	0.37	0.83	0.06	1.00	0.27	0.24	0.71
log(2.0)	1:1	0.30	1.00	0.40	0.36	0.32	0.20	1.00	0.26	0.23	0.20
	2:1	0.21	1.00	0.31	0.28	0.38	0.12	1.00	0.16	0.15	0.15
	4:1	0.14	1.00	0.28	0.26	0.61	0.06	1.00	0.12	0.11	0.35

We consider $P_G = P_E = 0.3$, $OR_{10} = OR_{01} = 1$ in all settings.
CC, Case-Control; CO, case-only; EB, empirical Bayes; TS, two-step.

constraint of gene-environment independence in the underlying population and yet can adapt itself to protect against large inflation of Type I error when the gene-environment independence assumption is violated.

Some specific results merit further discussion. Our studies in fixed parameter settings (Table IV) suggest that the EB procedure may not be able to maintain a desired Type I-error level exactly in all different scenarios of gene-environment dependence. Thus, if strict control of Type I error in all different scenarios of gene-environment dependence is used as the primary criterion for evaluating the methods, then standard case-control analysis remains the best option for analysis of case-control data in general. We, however, find it encouraging that when the violation of gene-environment independence is small, e.g. $\exp(\theta_{GE}) = 1.1$, then the EB procedure can maintain the Type I error at the nominal level and yet can gain substantial power over the standard case-control analysis. Empirical studies suggest that violation of gene-environment independence, when it occurs, would likely to be modest in most situations [Liu et al., 2004]. Moreover, the EB procedure is unbiased asymptotically. Thus, as sample size increases, any inflation of Type I error eventually disappears irrespective of the magnitude of gene-environment dependence.

We find it most interesting to compare alternative methods in terms of their Type I error and power after averaging them over likely scenarios of gene-environment dependence. The scenario presented in Table VI, for example, may correspond to a study where a large variety of genotype-exposure combinations are being studied, and for 80% of those combinations the gene-environment independence assumption is satisfied and for the rest the genotype-exposure log-odds-ratios have a distribution with 95% mass within ±log(1.5) limits. We find that under this kind of distribution, which we believe represents reasonable departure from gene-environment independence for large-scale association study, such as a genome-wide scan, the EB procedure, unlike the case-only and two-stage methods, on average can maintain the Type I error at a desired nominal level and yet can gain major power over the standard case-control analysis.

In this article, we have focussed on tests for multiplicative interactions. It is, however, important to recognize that the value of studying genetic and environmental exposures together does not necessarily stem from the ability to test for statistical interactions. Various alternative parameters, such as the joint effect of two exposures or the sub-group effects of one exposure within strata defined by the other exposure, may be useful for developing powerful test of association, understanding the public health impact of the exposures, targeting intervention and risk prediction. The simple EB procedure described in this article can be extended to carry out inference regarding such alternative parameters of interest. In recent years, for example, omnibus tests that can simultaneously account for genetic main effects and gene-environment/gene-gene interactions have received attention as a powerful approach for detection of disease of susceptibility loci [Chatterjee et al., 2006; Kraft et al., 2007]. The EB procedure has been extended by MC beyond the 2×4 table to estimate all of the parameters of a general logistic regression model using the framework of Chatterjee and Carroll [2005]. The general EB procedure can be used for developing more powerful versions of such omnibus tests.

Our consideration of case-control designs with smaller sampling rate for the controls than the cases reveals some intriguing observations. As m, the number of controls per case, decreases, the power of EB procedure diminishes at a much slower rate than that for the standard case-control analysis. Thus, if an EB-type procedure is to be used for analysis of interaction from case-control studies, then one could use smaller sampling ratio for controls than the cases, e.g. $m = 0.5$ or 0.25, thus reducing the cost of the study without reducing the efficiency proportionately. Of course, if one could completely rely on the gene-environment independence assumption, one could test for interaction using cases only. But the advantage of

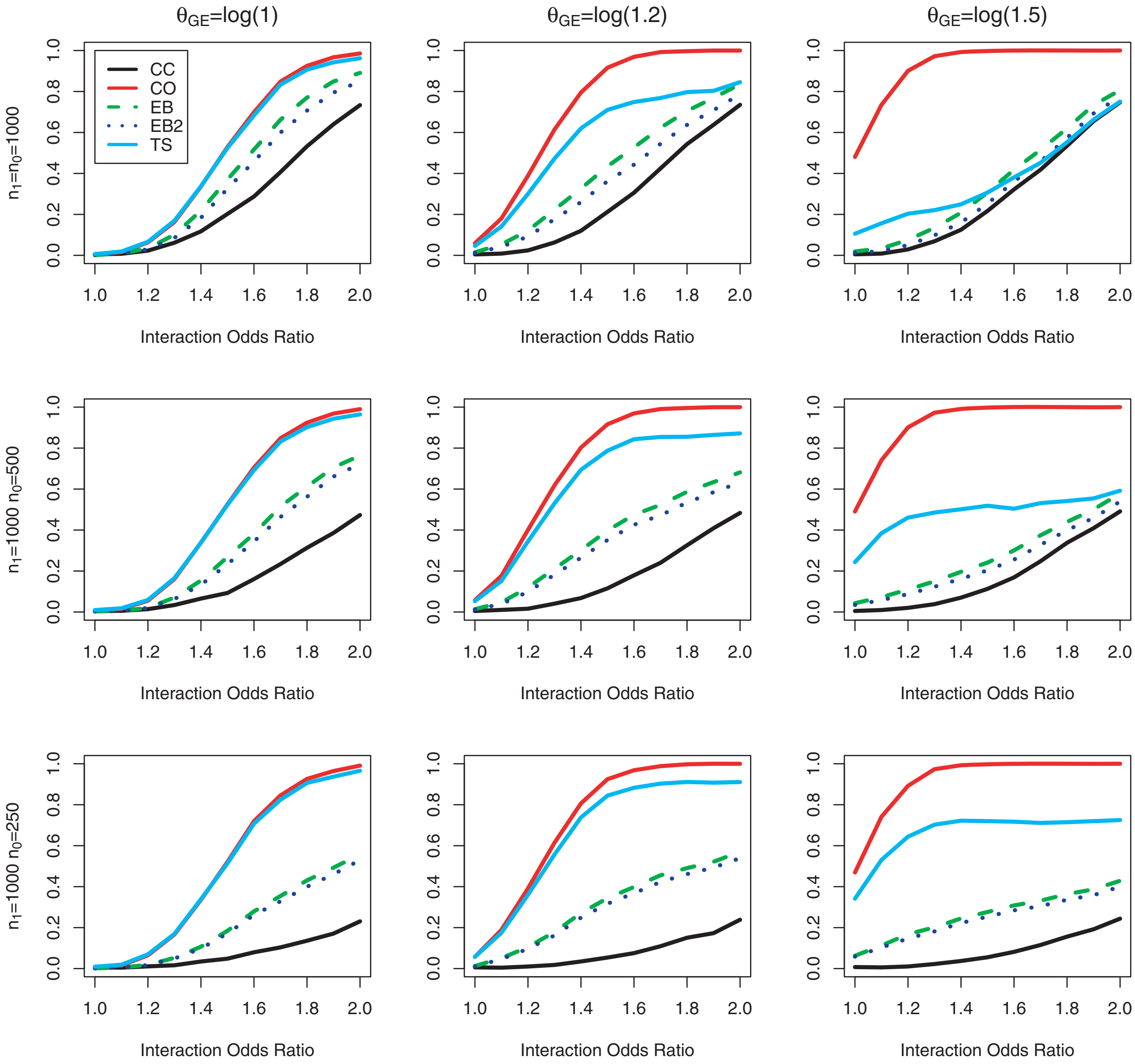

Fig. 3. Power for different estimators with 1,000 cases and number of controls varying across each row of the graphical array. The true value for θ_{GE} is set at 0, log(1.2) and log(1.5) across each column of the array. The horizontal axis in each plot represents the true value of the interaction odds ratio $\psi = \exp(\beta)$. The Type 1 error level is set at $\alpha = 0.005$. We consider $P_G = P_E = 0.3$, $\mathrm{OR}_{10} = \mathrm{OR}_{01} = 1$ in all settings. Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.

sampling controls is that they allow protection against false-positive results due to violation of the gene-environment independence assumption. It is promising that the EB procedure can provide good control of Type I error even with fairly small number of controls.

Implementation of such designs, however, requires careful evaluation of the main goals of a study. Epidemiologic researchers often design their studies with the goal of describing the association of a disease with certain types of exposures, such as genetic susceptibility, by itself and stratified by some other factors, such as an environmental exposure. In such situations, it is important to note that the strategy of sampling controls at a smaller rate than the cases although could be cost efficient for studying interactions, it would generally reduce the power of studies of main effect of the main exposure of interest. An alternative strategy that could be powerful and yet be cost-effective in such context would be to sample the controls at the same rate as the cases for the evaluation of the main exposure of interest, say the genetic markers, but limit any expensive ascertainment of the "other" type of exposure, such as biomarker-based evaluation of an environmental exposure, only on a smaller fraction of the controls.

Software: An Excel spreadsheet where one can input the cell frequencies of a 2 × 4 table to compute the five

TABLE VI. Integrated Type I error and power with varying number of controls under two settings (a) with $n_1 = 500, \alpha = 0.05$ and (b) $n_1 = 1,000, \alpha = 0.005$

α	n_0	β	CC	CO	EB	EB2	TS
0.05		log(1)	0.050	0.070	0.042	0.039	0.072
	500	log(1.25)	0.112	0.210	0.147	0.137	0.211
		log(1.5)	0.289	0.528	0.408	0.376	0.522
		log(2.0)	0.684	0.931	0.826	0.785	0.905
		log(1)	0.054	0.070	0.035	0.035	0.086
	250	log(1.25)	0.088	0.210	0.117	0.110	0.218
		log(1.5)	0.197	0.528	0.321	0.299	0.522
		log(2.0)	0.493	0.931	0.731	0.693	0.903
		log(1)	0.044	0.070	0.031	0.031	0.095
	125	log(1.25)	0.073	0.210	0.082	0.078	0.235
		log(1.5)	0.130	0.528	0.234	0.218	0.519
		log(2.0)	0.319	0.931	0.583	0.561	0.907
		log(1)	0.004	0.021	0.004	0.003	0.013
	1000	log(1.25)	0.040	0.133	0.065	0.055	0.118
		log(1.5)	0.204	0.524	0.356	0.313	0.510
		log(2.0)	0.726	0.969	0.873	0.835	0.946
		log(1)	0.006	0.021	0.005	0.005	0.020
0.005	500	log(1.25)	0.025	0.133	0.051	0.046	0.132
		log(1.5)	0.104	0.524	0.263	0.235	0.511
		log(2.0)	0.465	0.969	0.741	0.704	0.938
		log(1)	0.005	0.021	0.005	0.004	0.023
	250	log(1.25)	0.014	0.133	0.037	0.034	0.131
		log(1.5)	0.049	0.524	0.176	0.163	0.504
		log(2.0)	0.228	0.969	0.551	0.519	0.943

Type I errors (the rows corresponding to the null value of $\beta = log(1) = 0$) and powers are approximately integrated with respect to a mixture distribution for θ_{GE}, with 80% mass at 0 and 20% mass at $N(0, log(1.5)/2)$. The standard deviation parameter chosen such that roughly 95% of the θ_{GE} values fall within $\pm$log(1.5). We consider $P_G = P_E = 0.3$, $OR_{10} = OR_{01} = 1$ in all settings.

CC, Case-control; CO, case-only; EB, empirical Bayes; TS, two-step.

estimators we studied, the corresponding standard errors, P-values and CI is available at http://www.sph.umich.edu/bhramar/public_html/research. The R-codes for simulating power for the different tests of interaction under general study settings is also available in the above web site. The matlab software for the general EB procedure for estimating all the parameters of a logistic regression model is available at http://dceg.cancer.gov/about/staff-bios/chatterjee-nilanjan♯software.

ACKNOWLEDGMENTS

The research of Bhramar Mukherjee was partially supported by NSF young investigator grant DMS 07-06935 and NIH grant R03 CA130045-01. The Molecular Epidemiology of Colorectal Cancer study is supported via NIH grant R01 CA81488. Victor Moreno was funded by the Spanish Secretaria de Estado de Universidades e Investigacion, PR2006-0474 (V.M.). The research of Dr. Nilanjan Chatterjee was supported by a Gene-Environment Initiative (GEI) grant from the National Heart Lung and Blood Institute (R01 HL091172-01) and by the Intramural Research Program of the National Cancer Institute.

REFERENCES

Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. 2001. Limitations of the case-only design for identifying gene-environment interactions. Am J Epidemiol 154:687–693.

Barrett JH, Smith G, Waxman R, Gooderham N, Lightfoot T, Garner RC, Augustsson K, Wolf CR, Bishop DT, Forman D. 2003. Investigation of interaction between **N**-acetyltransferase 2 and heterocyclic amines as potential risk factors for colorectal cancer. Carcinogenesis 24:275–282.

Breslow NE, Day NE. 1987. Statistical Methods in Cancer Research. Vol. II: Design and Analysis of Cohort Studies. Lyon: IARC.

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

Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. 2006. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am J Hum Genet 79:1002–1016.

Greenland S. 1993. Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum likelihood, preliminary-testing, and empirical Bayes regression. Stat Med 12:717–736.

Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. 2007. Exploiting gene-environment interaction to detect genetic associations. Hum Hered 63:111–119.

Lin HJ, Probst-Hensch NM, Louie AD, Kau IH, Witte JS, Ingles SA, Frankl HD, Lee ER, Haile RW. 1998. Glutathione transferase null genotype, broccoli, and lower prevalence of colorectal adenomas. Cancer Epidemiol Biomarkers Prev 8:647–652.

Liu X, Fallin MD, Kao WH. 2004. Genetic dissection methods: designs used for tests of gene-environment interaction. Curr Opin Genet Dev 14:241–245.

Moslehi R, Chatterjee N, Church TR, Chen J, Yeager M, Weissfield J, Hein DW, Hayes RB. 2006. Cigarette smoking *n*-acetyltransferase genes and the risk of advanced colorectal adenoma. Pharmacogenomics 7:819–829.

Mukherjee B, Chatterjee N. 2007. Exploiting gene-environment independence for analysis of case-control studies: an empirical-Bayes type shrinkage estimator to trade off between bias and efficiency. Biometrics: in press. [E-pub doi: 10.111/j.1541-0420.2007.00953.x].

Piegorsch WW, Weinberg CR, Taylor J. 1994. Non hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 13:153–162.

Poynter JN, Gruber SB, Higgins PD, Almog R, Bonner JD, Rennert HS, Low M, Greenson JK, Rennert G. 2005. Statins and the risk of colorectal cancer. N Engl J Med 352:2184–2192.

Roberts-Thomson IC, Ryan P, Khoo KK, Hart WJ, McMichael AJ, Butler RN. 1996. Diet, acetylator phenotype and risk of colorectal neoplasia. Lancet 347:1372–1374.

Satten GA, Kupper LL. 1993. Inferences about exposure-disease associations using probability-of-exposure information. J Am Sta Assoc 88:200–208.

Smits KM, Gaspari L, Weijenberg MP, Dolzan V, Golka K, Roemer HC, Nedelcheva Kristensen V, Lechner MC, Mehling GI, Seidegard J, Strange RC, Taioli E. 2003. Interaction between smoking, GSTM1 deletion and colorectal cancer: results from the GSEC study. Biomarkers 8:299–310.

Umbach DM, Weinberg CR. 1997. Designing and analysing case-control studies to exploit independence of genotype and exposure. Stat Med 16:1731–1743.

Willett WC, Reynolds RD, Cottrell-Hoehner S, Sampson L, Browne ML. 1987. Validation of a semi-quantitative food frequency questionnaire: comparison with a 1-year diet record. J Am Diet Assoc 87:43–47.