Biometrics 64, 685–694
September 2008
DOI: 10.1111/j.1541-0420.2007.00953.x

Exploiting Gene-Environment Independence for Analysis of Case–Control Studies: An Empirical Bayes-Type Shrinkage Estimator to Trade-Off between Bias and Efficiency

Bhramar Mukherjee[1] and Nilanjan Chatterjee[2]

[1]Department of Biostatistics, University of Michigan,
Ann Arbor, Michigan 48109, U.S.A.
email: bhramar@umich.edu
[2]Division of Cancer Epidemiology and Genetics, National Cancer Institute,
National Institutes of Health, Department of Health and Human Services,
Rockville, Maryland 20852, U.S.A.
email: chattern@mail.nih.gov

Summary. Standard prospective logistic regression analysis of case–control data often leads to very imprecise estimates of gene-environment interactions due to small numbers of cases or controls in cells of crossing genotype and exposure. In contrast, under the assumption of gene-environment independence, modern "retrospective" methods, including the "case-only" approach, can estimate the interaction parameters much more precisely, but they can be seriously biased when the underlying assumption of gene-environment independence is violated. In this article, we propose a novel empirical Bayes-type shrinkage estimator to analyze case–control data that can relax the gene-environment independence assumption in a data-adaptive fashion. In the special case, involving a binary gene and a binary exposure, the method leads to an estimator of the interaction log odds ratio parameter in a simple closed form that corresponds to an weighted average of the standard case-only and case–control estimators. We also describe a general approach for deriving the new shrinkage estimator and its variance within the retrospective maximum-likelihood framework developed by Chatterjee and Carroll (2005, *Biometrika* **92,** 399–418). Both simulated and real data examples suggest that the proposed estimator strikes a balance between bias and efficiency depending on the true nature of the gene-environment association and the sample size for a given study.

Key words: Case-only designs; Gene-environment interaction; Profile likelihood; Retrospective analysis; Semiparametrics.

1. Introduction

While prospective logistic regression remains an established method to analyze case–control data, recent problems emerging in genetic epidemiology have attracted attention to retrospective analysis because it can incorporate certain scientifically plausible constraints on the exposure distribution in the underlying population. In studies of gene-environment association with disease, for example, it often may be realistic to assume that genetic susceptibilities (G) and environmental exposures (E) are independent of each other in the underlying population. Piegorsch, Weinberg, and Taylor (1994) noticed that under G-E independence and assuming a rare disease, the interaction odds ratio between G and E can be estimated using the association odds ratio between these factors in cases alone. Moreover, this "case-only" estimate of interaction can be much more precise than that obtained from standard case–control analysis. Umbach and Weinberg (1997) generalized this idea to show that the maximum-likelihood estimates (MLEs) of all of the parameters of a logistic regression model involving categorical exposures can be obtained under the independence assumption by fitting a suitably constrained log-linear model to the case–control data. Recently, Chatterjee and Carroll (2005) developed a rigorous semiparametric framework for retrospective maximum-likelihood (ML) analysis of case–control data under the gene-environment independence assumption in a general setting that may involve continuous exposures, nonrare diseases, and population stratification. The classical result about the equivalence of prospective and retrospective maximum likelihood (Andersen, 1970; Prentice and Pyke, 1979), which assumes unconstrained covariate distribution, does not hold in this setting and the retrospective approach is generally more efficient (Chatterjee and Carroll, 2005). Similar gain in efficiency has been also noted for retrospective methods that can incorporate constraints on the genotype distribution imposed by population genetic laws such as Hardy–Weinberg equilibrium (HWE) (Epstein and Satten, 2003; Satten and Epstein 2004; Spinka, Carroll, and Chatterjee, 2005; Lin and Zeng, 2006; Chen and Chatterjee, 2007).

A major hindrance for practical use of retrospective methods, in spite of their efficiency advantage, has been the potential for large bias in these methods when some of the

underlying assumptions such as gene-environment independence or HWE are violated (Albert et al., 2001; Satten and Epstein, 2004; Chatterjee and Carroll, 2005; Spinka et al., 2005). A number of alternative strategies for relaxing the underlying assumptions have been proposed. Chatterjee and Carroll (2005) considered a model that can account for gene-environment dependence due to population stratification. Satten and Epstein (2004) and Lin and Zeng (2006) considered relaxing the HWE assumption based on alternative, more flexible population genetics models. These models alleviate the concern of bias somewhat, but may not be adequate because they only capture certain types of departures from the underlying constraints. One could also use a two-stage procedure where, at first, one formally tests for the adequacy of the underlying assumption(s) based on the data itself and then uses the outcome of that test to decide whether to use the efficient retrospective or the more robust prospective method for odds ratio estimation. For a given study of modest sample size, however, the power of the tests for HWE or/and 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 much higher type I error level than desired (Albert et al., 2001).

In this article, we propose a novel solution to the bias versus efficiency dilemma of retrospective methods using a simple stochastic framework that allows for uncertainty around the assumption of gene-environment independence. We show how the magnitude of the uncertainty parameter can be estimated from the data itself. We then use this estimate of the uncertainty parameter in an empirical Bayes (EB) fashion to obtain a shrinkage estimator that "shrinks" the MLEs of disease odds ratio parameters under a general model for *G-E* dependence to those obtained under the assumption of *G-E* independence.

In Section 2, we consider a simple scenario involving a binary G and a binary E, where the proposed estimator of the interaction odds ratio can be derived in the form of a simple weighted average of the standard "case-only" and "case–control" estimators. Simulation studies show that in finite samples, the proposed estimator can strike a balance between bias and efficiency depending on the changing scenarios of gene-environment association. Motivated by these results, in Section 3, we then describe a general approach for deriving such shrinkage estimators for all of the parameters of a general logistic regression model. We consider the retrospective maximum likelihood framework developed by Chatterjee and Carroll (2005), but relax the underlying gene-environment independence assumption by modeling the gene frequencies as a function of the environmental covariates using a logistic or polytomous-logistic regression model with random coefficients. We then develop a general theory for constructing an EB-type shrinkage estimator based on the profile likelihood of the data that avoids estimation of the high-dimensional nuisance parameters associated with the marginal distribution of the environmental covariates. Further simulation studies are conducted to investigate the performance of the general estimator when there are two environmental exposures, one of which is associated with G and the other is not. In both Sections 2 and 3, a method for variance estimation for the respective EB estimators is proposed. In Section 4, we analyze two datasets, both providing evidence of how the EB estimate is tracking the MLEs from the constrained or unconstrained model depending upon the strength of *G-E* association in the respective studies. Section 5 presents discussion and possibilities for future work. Additional simulation results and technical details are relegated to supplementary Web Appendices.

Table 1
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

2. Binary Genetic and Environmental Factors

In this section, we consider the simple set-up of an unmatched case–control study with a binary genetic factor G and a binary environmental exposure E. Let $E=1$ ($E=0$) denote an exposed (unexposed) individual and $G=1$ ($G=0$) denote whether an individual is a carrier (noncarrier) of the susceptible 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 1.

Let r_{dge} and p_{dge} denote the observed cell count and the unknown true cell probability, respectively, for the configuration $D=d$, $G=g$ and $E=e$, d,g,$e=0, 1$. Let $\mathbf{r}_0 = (r_{000}, r_{001}, r_{010}, r_{011})$ 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. Let $\boldsymbol{p}_0 = (p_{000}, p_{001}, p_{010}, p_{011} = 1 - p_{000} - p_{001} - p_{010})$ 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, $\boldsymbol{r}_0 \sim$ Multinomial $(\mathrm{n}_0, \boldsymbol{p}_0)$ and $\boldsymbol{r}_1 \sim$ Multinomial $(\mathrm{n}_1, \boldsymbol{p}_1)$. Let $OR_{10} = p_{000}\, p_{101}/p_{001}\, p_{100}$ denote the odds ratio associated with E for nonsusceptible subjects ($G=0$), $OR_{01} = p_{000}\, p_{110}/p_{010}\, p_{100}$ denote the odds ratio associated with G for unexposed subjects ($E=0$), and $OR_{11} = p_{000}\, p_{111}/p_{011}\, p_{100}$ denote the odds ratio associated with $G=1$ and $E=1$ compared to the baseline category $G=0$ and $E=0$. Therefore, $\psi = OR_{11}/(OR_{10}OR_{01}) = (p_{001}p_{010}p_{100}p_{111})/(p_{000}p_{011}p_{101}p_{110})$ is the multiplicative interaction parameter of interest.

To this end, let us consider a measure of *G-E* association in the control population, namely,

$$\theta_{GE} = \log\{(p_{000}p_{011})/(p_{001}p_{010})\}. \quad (1)$$

The assumption of *G-E* independence, together with the rare disease approximation implies $\theta_{GE} = 0$ (Schmidt and Schaid, 1999). When one is not certain about the *G-E* independence, one may conceptually posit a stochastic framework for the underlying true parameter θ_{GE} as,

Table 2

Simulation results showing MSE and bias (in parentheses) in estimation of the interaction parameter $\beta = log(\psi)$ for different methods under varying scenarios of G-E association. The value of θ_{GE} is the control odds ratio between G and E. The prevalences of G and E were fixed at $P_G = P_E = 0.3$ in the control population. The parameters in the disease risk model were set at $OR_{10} = OR_{01} = 1$ and $\beta = log\ (\psi) = log\ (2) = 0.6931$. Results are based on 5000 simulated datasets.

		Sample size		
		$n_0 = n_1 = 100$	$n_0 = n_1 = 200$	$n_0 = n_1 = 500$
$\theta_{GE} = 0$	Case–control	0.46 (0.03)	0.22 (0.02)	0.08 (0.00)
	Case-only	0.20 (0.01)	0.10 (0.01)	0.04 (0.00)
	EB	0.29 (0.02)	0.14 (0.01)	0.05 (0.00)
	EB^{+a}	0.27 (0.01)	0.13 (0.01)	0.05 (0.00)
	Two-stage	0.26 (0.00)	0.13 (0.01)	0.05 (0.00)
	IL^{b}	0.28 (0.02)	0.14 (0.01)	0.07 (0.01)
$\theta_{GE} = \log\ (1.25)$	Case–control	0.45 (0.02)	0.21 (0.00)	0.08 (0.01)
	Case-only	0.26 (0.24)	0.15 (0.23)	0.09 (0.23)
	EB	0.31 (0.12)	0.16 (0.10)	0.07 (0.09)
	EB^{+}	0.30 (0.16)	0.16 (0.14)	0.08 (0.12)
	Two-stage	0.31 (0.16)	0.19 (0.15)	0.10 (0.13)
	IL	0.35 (0.16)	0.17 (0.10)	0.09 (0.06)
$\theta_{GE} = \log\ (1.5)$	Case–control	0.45 (0.02)	0.21 (0.01)	0.08 (0.00)
	Case-only	0.39 (0.43)	0.27 (0.42)	0.21 (0.41)
	EB	0.37 (0.19)	0.20 (0.16)	0.10 (0.12)
	EB^{+}	0.36 (0.25)	0.21 (0.21)	0.11 (0.15)
	Two-stage	0.44 (0.27)	0.28 (0.22)	0.15 (0.13)
	IL	0.38 (0.19)	0.24 (0.17)	0.10 (0.07)
$\theta_{GE} = \log\ (2)$	Case–control	0.45 (0.03)	0.21 (0.02)	0.08 (0.01)
	Case-only	0.74 (0.73)	0.60 (0.71)	0.54 (0.70)
	EB	0.46 (0.27)	0.25 (0.20)	0.10 (0.11)
	EB^{+}	0.50 (0.33)	0.28 (0.24)	0.11 (0.12)
	Two-stage	0.67 (0.34)	0.38 (0.19)	0.11 (0.03)
	IL	0.53 (0.25)	0.28 (0.13)	0.10 (0.03)

[a]The EB estimator using the positive part variance estimator $\hat{\tau}_+^2 = \max(0, \hat{\theta}_{GE}^2 - \hat{\sigma}_{\hat{\theta}_{GE}}^2)$.

[b]The MLE of the interaction parameter as obtained from maximizing the IL. The full likelihood for the case–control data was integrated with respect to the stochastic parameter θ_{GE} with density $N(0, \tau^2)$. The IL was approximated by a 30 point Gauss–Hermite quadrature and then maximized over τ^2 and all other model parameters.

$\theta_{GE} \sim N(0,\ \tau^2)$, where τ^2 reflects a measure of uncertainty about the independence assumption.

Next we investigate how one can estimate the prior variability τ^2 using the data itself.

The MLE of the *G-E* odds ratio among controls, namely, θ_{GE}, is given by

$$\hat{\theta}_{GE} = \log\left\{(r_{000}r_{011})/(r_{001}r_{010})\right\}.$$

Standard likelihood theory implies that, given $\theta_{GE}, \hat{\theta}_{GE} \sim N(\theta_{GE}, \sigma^2_{\hat{\theta}_{GE}})$, where an estimate of the asymptotic variance is given by $\hat{\sigma}^2_{\hat{\theta}_{GE}} = \Sigma^1_{g=0}\Sigma^1_{e=0}(1/r_{0ge})$. Marginalizing over θ_{GE}, it follows that marginally $\hat{\theta}_{GE} \sim N(0, \tau^2 + \sigma^2_{\hat{\theta}_{GE}})$. Thus, based on the marginal variance of $\hat{\theta}_{GE}$, a consistent estimator of the unknown hyperparameter τ^2 can be obtained simply as (Morris, 1983; Greenland, 1993), $\hat{\tau}_+^2 = \max(0, \hat{\theta}^2_{GE} - \hat{\sigma}^2_{\hat{\theta}_{GE}})$. We consider a more conservative estimate of the prior variance obtained as $\hat{\tau}^2 = \hat{\theta}^2_{GE}$ because it leads to a convenient form for the variance expression of our subsequently proposed estimator of $\beta = \log\ (\psi)$. Simulation studies show there is essentially no loss of efficiency using $\hat{\tau}^2$ instead of $\hat{\tau}_+^2$ (see Table 2).

With the above stochastic framework in mind, we now propose a new shrinkage estimator by combining two commonly used estimators of $\log\ (\psi) = \beta$, the one obtained from using case–control data ($\hat{\beta}_{CC}$), and the other obtained from cases alone ($\hat{\beta}_{CO}$), with the corresponding formulae given by

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

Note that $\hat{\beta}_{CC}$ is the unconstrained MLE of β given the data shown in Table 1, whereas $\hat{\beta}_{CO}$ is the MLE under the constraint of *G-E* independence, i.e., $\theta_{GE} = 0$ (Umbach and Weinberg, 1997). Let $\hat{\sigma}^2_{CC} = \Sigma^1_{d=0}\Sigma^1_{g=0}\Sigma^1_{e=0}1/r_{dge}$ denote the estimated asymptotic variance of the case–control estimator $\hat{\beta}_{CC}$. We propose the following weighted estimator of the interaction parameter:

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

We observe $\hat{\beta}_{EB}$ can be viewed as a shrinkage estimator where $\hat{\beta}_{CC}$, the robust case–control estimator, has been shrunk toward $\hat{\beta}_{CO}$, the possibly efficient estimator of β under the

assumption of *G-E* independence. The specific form of the "shrinkage" weights is motivated by the EB perspective (Morris, 1983; Greenland, 1993). The EB structure resembles the form of a posterior mean obtained in a classical Bayesian analysis under a normal-normal model (Berger, 1985, p. 131), with the prior variance substituted by an estimate obtained using a method of moments approach. Further justification of the proposed estimator as a special case of a more general framework is provided in Section 3. We, however, recognize that this estimator is not a true "Bayes" or "empirical Bayes" estimator in a strict technical sense as we are not carrying out a proper full Bayesian analysis here with a joint prior structure on all the parameters of interest; we are using prior structure *only* on the "nuisance parameter" θ_{GE} and embedding that prior uncertainty in the estimation paradigm for the parameter of interest β. In this sense, the proposed method has a conceptual resemblance to the partial Bayes inference introduced by Cox (1975).

We observe that as $\hat{\tau}^2 = \hat{\theta}^2_{GE} \to 0$, i.e., as the data provide evidence in favor of *G-E* independence in the control population, $\hat{\beta}_{EB} \to \hat{\beta}_{CO}$, and as $\hat{\tau}^2 = \hat{\theta}^2_{GE} \to \infty$, i.e., as the uncertainty regarding *G-E* independence in control population becomes stronger, $\hat{\beta}_{EB} \to \hat{\beta}_{CC}$. Because $\hat{\beta}_{CC} = \hat{\beta}_{CO} - \hat{\theta}_{GE}$, one can also express the estimator in (2) as

$$\hat{\beta}_{EB} = \hat{\beta}_{CO} - K(\hat{\sigma}^2_{CC}, \hat{\tau})\hat{\theta}_{GE}, \tag{3}$$

where the shrinkage factor $K(\hat{\sigma}^2_{CC}, \hat{\tau}) = \{1 + (\hat{\sigma}^2_{CC}/\hat{\tau}^2)\}^{-1}$ "shrinks" $\hat{\theta}_{GE}$, the control log odds ratio between *G* and *E*, to its hypothesized mean value of zero under the *G-E* independence assumption.

In the following subsection, we study the performance of the proposed estimator relative to a number of alternative estimators under varying scenarios of *G-E* association.

2.1 *Simulation Study for the 2 × 4 Table*

Although the estimate $\hat{\beta}_{EB}$ is postulated in a Bayesian framework, it is purely a functional of the data (namely, the multinomial counts, $\mathbf{r}_0$ and $\mathbf{r}_1$). The implicit background of assuming a normal prior with variance τ^2 does not play any explicit role in the computation of this estimator. Thus, in our simulation, we first study the finite sample properties of this estimator in the standard fixed parameter setting of a frequentist paradigm. A second set of simulations where θ_{GE} is generated from a random distribution is contained in Web Appendix A (Web Table 1).

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 $\mathbf{p}_0$ by solving a system of equations.

We then set the values of OR_{10}, OR_{01}, and ψ, which together with $\mathbf{p}_0$, 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 and compute the case–control, case-only, and the proposed shrinkage estimator under varying scenarios. We also include the two-stage estimator proposed by Albert et al. (2001) in our simulation study. The two-stage estimator first tests for *G-E* independence in controls by testing the hypothesis H_0: $\theta_{GE} = 0$ at a significance level of $\alpha = 0.05$, and based on the acceptance/rejection of this hypothesis, the case-only or the case–control estimator is then used. Based on the suggestion of an anonymous reviewer, in our simulation we also considered an alternative approach where all of the disease odds ratio parameters and τ^2 are jointly estimated by maximization of the "integrated likelihood" (IL) obtained by integration of the standard case–control likelihood with respect to the stochastic parameter θ_{GE}.

Table 2 presents the mean-squared error (MSE) and bias of different estimators of the interaction parameter $\beta = \log(\psi)$, when $P_G = P_E = 0.3$ and $OR_{10} = OR_{01} = 1$. The *G-E* odds ratio among controls, namely, $\exp(\theta_{GE})$ is varied at four different values, 1, 1.25, 1.5, and 2. The true value of β is set at $\log(2)$. The results are based on 10,000 simulated datasets. The results clearly indicate that the proposed EB estimator follows the case–control and the case-only estimators based on the value of θ_{GE} in a data-adaptive way. It has much reduced bias and MSE compared to the case-only estimator under violation of the independence assumption. It also maintains significantly smaller MSE compared to the case–control estimator under independence as well as under modest departures from independence. Under large departures from independence, the EB estimator performs very comparably to the case–control estimator. In contrast, the performance of the case-only estimator deteriorates sharply as one moves away from the independence assumption. Unlike the case-only estimator, which is asymptotically biased, any residual bias in the EB estimator goes to zero in large samples. The EB$^+$ estimator, which uses the estimate of $\hat{\tau}^2_+$ instead of $\hat{\tau}^2$, performs comparably as EB in terms of MSE, but has somewhat larger bias under departures from the independence assumption. The two-stage estimator does not perform well in terms of bias and MSE, especially in small samples. The IL method, which can be computationally intensive, performs very similar to the simpler EB estimator in terms of MSE. Interestingly, however, under departure from gene-environment independence, the bias of IL seems to go to zero at a faster rate than EB as sample size increases.

Consistency of the proposed estimator: The proposed estimator is $n^{\frac{1}{2}}$ consistent in the fixed parameter frequentist setting. In particular, when gene-environment independence is violated, i.e., $\theta_{GE} \neq 0$, it is easy to see that as $n \to \infty$ and hence $\hat{\sigma}^2_{CC} \to 0$, the EB estimator will converge to the case–control estimator at a $n^{\frac{1}{2}}$ rate. For $\theta_{GE} = 0$, the consistency of the EB estimator can be seen from the representation (3) and noting that the case-only estimator $\hat{\beta}_{CO}$ in this case is $n^{\frac{1}{2}}$ consistent and the second term converges to a zero mean random variable, also at a $n^{\frac{1}{2}}$ rate.

Variance of the proposed estimator: In the following, we propose a method to obtain an asymptotic variance expression for $\hat{\beta}_{EB}$. Because $\hat{\sigma}^2_{CC} \to 0$ at the rate of $O(1/n)$, one may ignore the variation in $\hat{\sigma}^2_{CC}$ and treat this as a constant while obtaining the first-order $n^{\frac{1}{2}}$-asymptotic approximation of the proposed estimator. Under this setting, the first and second term in (3) could be considered as asymptotically independent as the first term depends only on cases, and the second depends only on controls. Using Taylor's expansion on the second term, considering it as a function of $\hat{\tau} = \hat{\theta}_{GE}$, and

treating $\hat{\sigma}^2_{CC}$ as a constant, we have an estimator of variance of the form,

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

This estimate of the variance in (4), namely $\widehat{V}_A$, performs remarkably well even in small samples ($n_0 = n_1 = 100$) when compared to the empirical variance (see Web Appendix A, Web Table 2).

For constructing interval estimates, we used Wald-type confidence intervals for the log odds ratio parameters based on the standard errors derived from the above formula. The approximate normality of the proposed estimator even with smaller sample sizes can be seen in Web Figure 1 in the supplementary online material. Coverage probabilities for such Wald-type confidence intervals are furnished in Web Table 2.

3. The General Case: Profile Likelihood and Empirical Bayes

Chatterjee and Carroll (2005) have described a general approach for estimation of the parameters of a logistic regression model from case–control studies under the assumption of gene-environment independence. They allowed for the presence of stratification factors ($\boldsymbol{S}$) such as ethnicity which could be related to both G and E. They consider the following factorization of the retrospective likelihood,

$$\begin{aligned} L^R &= \text{pr}(G, E, \boldsymbol{S} \mid D) \\ &= \frac{\text{pr}(D \mid G, E, \boldsymbol{S})\text{pr}(G \mid E, \boldsymbol{S})\text{pr}(E, \boldsymbol{S})}{\sum_{\text{G,E,s}} \text{pr}(D \mid G, E, \boldsymbol{S})\text{pr}(G \mid E, \boldsymbol{S})\text{pr}(E, \boldsymbol{S})}. \end{aligned} \quad (5)$$

For continuous exposure E, the sum with respect to E in the denominator of (5) is replaced by an integral. The ingredients of the retrospective likelihood are constituted in the following way. Assume a logistic disease incidence model pr(D = 1 | G, E, $\boldsymbol{S}$) = $H\{\gamma_0 + m(G, E, \boldsymbol{S}\ ; \gamma_1)\}$, where $H(u) = (1 + \exp(-u))^{-1}$ and $m(\cdot)$ is a known but arbitrary function. The joint distribution function for (E, $\boldsymbol{S}$) is allowed to remain completely unrestricted (nonparametric). Under *G-E* independence, conditional on $\boldsymbol{S}$, pr(G | E, $\boldsymbol{S}$) = pr (G | $\boldsymbol{S}$). Assuming a binary genetic factor G, consider a logistic model of the form

$$\text{pr}(G = 1 \mid E, \boldsymbol{S}) = H\{\eta_0 + \eta_1\, \boldsymbol{S}\}. \quad (6)$$

The model could be extended to a multinomial logistic model for a general categorical G, such as genotype data for single nucleotide polymorphisms which is typically coded as 0, 1 or 2 by counting the number of variant alleles carried by an individual. We will refer to (6) as the independence model, or the constrained model. To relax the assumption of *G-E* independence, one can expand the model in (6) to

$$\text{pr}(G = 1 \mid E, \boldsymbol{S}) = H\{\eta_0 + \eta_1\, \boldsymbol{S} + \theta E\}, \quad (7)$$

where θ is a measure of dependence between G and E. We will refer to (7) as the dependence or unconstrained model. Clearly, (6) can be viewed as a special case of (7) with $\theta = 0$.

The MLEs for the parameters $\omega = (\gamma, \eta)$ under model (6) as well as those for $\omega = (\gamma, \eta, \theta)$ under model (7) can be obtained using the profile-likelihood techniques of Chatterjee and Carroll (2005). In particular, the estimates of the ω-parameters that would maximize the retrospective likelihood L^R, while allowing the distribution of $Z = (E, \boldsymbol{S})$ to remain completely nonparametric, can be obtained by maximizing a simpler pseudolikelihood of the form $L^* = \text{pr}$ (D, G | E, $\boldsymbol{S}$, $R = 1$), where the conditioning event $R = 1$ reflects the outcome dependent sampling mechanism for case–control studies. Computationally, the likelihood L^* is much more tractable as it does not require estimation of the high-dimensional "nuisance parameters" involved in specification of the distribution of Z. The details of the estimation method are provided in Chatterjee and Carroll (2005), and we use their developed software to implement the two models. In the following, the MLE for the common set of regression parameters $\beta = (\gamma, \eta)$ under the unconstrained and constrained models will be denoted by $\hat{\beta}_{ML}$ and $\hat{\beta}^0_{ML}$, respectively.

Before we proceed to form the EB-type shrinkage estimator for this particular context, we consider a general framework where one is interested in estimating a set of focus parameters β in the presence of prior information on a set of "nuisance" parameters θ. The general paradigm itself is a novel feature of this article.

Suppose $\zeta = (\beta, \theta)^T$ denotes a column vector of parameters, where β denotes a set of focus parameters and θ denotes a set of nuisance parameters. Let the dimensions of β and θ be p and m, respectively. Let $\zeta_0 = (\beta_0, \theta_0)^T$ denote the true values of the parameters in the population. Assume that one is willing to postulate a prior distribution for θ as $\text{MVN}_\text{m}(0, \boldsymbol{A})$, a m-dimensional zero-mean multivariate normal distribution with variance–covariance matrix $\boldsymbol{A}$. The goal is to conduct inference on β, without any further prior specification on β. Intuitively, given θ and in the absence of any prior information on β, a natural way to estimate β would be to use $\hat{\beta}_{ML}(\theta)$, the profile MLE of β for fixed θ. In the following, we show how to utilize the prior information on θ while working with the profile MLE $\hat{\beta}(\theta)$. Define $\beta(\theta)$ to be the limiting value of $\hat{\beta}_{ML}(\theta)$ which is a population parameter with $\beta(\theta) = \beta_0$ when θ is fixed at the true value θ_0. Note that the constrained MLE for β, with $\theta = 0$, can be written as $\hat{\beta}^0_{ML} = \hat{\beta}_{ML}(\theta = 0)$, and the unconstrained MLE can be written as $\hat{\beta}_{ML} = \hat{\beta}_{ML}(\theta = \hat{\theta}_{ML})$.

Let us then consider the general problem of EB estimation of a general vector function $\phi = f(\theta)$, of dimension p when the argument θ ($m \times 1$) has a prior $MVN_m(0, \boldsymbol{A})$. By applying Taylor's expansion around $\theta = 0$, the prior on ϕ could be linearly approximated as $\phi \sim MVN_p(f(0), \{f'(0)\}^\top A f'(0))$, where $f'(\theta) = \partial f^\top(\theta)/\partial\theta$ is the gradient matrix of dimension $m \times p$. Let $\hat{V}_\phi$ be the estimated asymptotic variance of $f(\hat{\theta}_{ML})$. Then an approximation to the Bayes estimate of $\phi = f(\theta)$ for a fixed $\boldsymbol{A}$ is given by

$$\begin{aligned} \hat{\phi} = \{f'(0)\}^\top \boldsymbol{A} f'(0) \left[\hat{V}_\phi + \{f'(0)\}^\top \boldsymbol{A}\, \{f'(0)\}\right]^{-1} f(\hat{\theta}_{ML}) \\ + \hat{V}_\phi \left[\hat{V}_\phi + \{f'(0)\}^\top \boldsymbol{A} f'(0)\right]^{-1} f(0). \end{aligned} \quad (8)$$

By applying (8), the Bayes estimator of $\beta = \beta(\theta)$ in our setting can be approximated for a known value of the prior covariance

matrix $\boldsymbol{A}$ as,

$$\widehat{\beta(\theta)} = \Delta^\top \boldsymbol{A}\Delta(\hat{V}_{\hat{\beta}_{ML}} + \Delta^\top \boldsymbol{A}\Delta)^{-1}\beta(\hat{\theta}_{ML}) + \hat{V}_{\hat{\beta}_{ML}}(\hat{V}_{\hat{\beta}_{ML}} + \Delta^\top \boldsymbol{A}\Delta)^{-1}\beta(0), \quad (9)$$

where $\Delta = \partial\beta^\top(\theta)/\partial\theta$ is the gradient matrix of dimension $m \times p$ evaluated at $\theta = 0$. Note that $\Delta^\top \boldsymbol{A}\Delta$ is a $p \times p$ matrix where p is the dimension of β. Now (9) itself cannot be used to estimate β as it involves the unknown function $\beta(\theta)$. We propose to plug in $\hat{\beta}_{ML}(\theta)$ for $\beta(\theta)$. Further, by observing the identity $\boldsymbol{S}_{\hat{\beta}_{ML}(\theta)}(\theta) \equiv 0$, where $\boldsymbol{S}_\beta(\theta)$ denotes the ML-score function for β given θ, by chain rule of derivatives, one can derive an estimate of Δ as

$$\hat{\Delta} = \frac{\partial\hat{\beta}_{ML}^\top}{\partial\theta}(\theta = 0) = -I_{\theta\beta}(\theta = 0)\{I_{\beta\beta}(0)\}^{-1}. \quad (10)$$

Here $I_{\theta\beta}$ and $I_{\beta\beta}$ denote suitable information matrices under the unconstrained model. In alignment with the EB spirit, we now estimate the prior hyperparameter $\boldsymbol{A}$ by a conservative upper bound to its marginal MLE, given by $\hat{\theta}_{ML}\hat{\theta}_{ML}^\top$. Thus the final form of our proposed estimate is given by

$$\hat{\beta}_{EB} = \hat{\Delta}^\top\hat{\theta}_{ML}\hat{\theta}_{ML}^\top\hat{\Delta}\left(\hat{V}_{\hat{\beta}_{ML}} + \hat{\Delta}^\top\hat{\theta}_{ML}\hat{\theta}_{ML}^\top\hat{\Delta}\right)^{-1}\hat{\beta}_{ML} + \hat{V}_{\hat{\beta}_{ML}}\left(\hat{V}_{\hat{\beta}_{ML}} + \hat{\Delta}^\top\hat{\theta}_{ML}\hat{\theta}_{ML}^\top\hat{\Delta}\right)^{-1}\hat{\beta}^0_{ML}. \quad (11)$$

Computationally, this requires only fitting the constrained model and the unconstrained model and evaluating the variance covariance components for the unconstrained model at $\theta = 0$. We note that in the above calculations a key step is to use the first-order Taylor's expansion to approximate the variance of the function $\beta(\theta)$. In Section 5, we discuss potential limitations of this approximation and some associated remedies.

Revisiting the 2 × 4 case: Now consider our proposed estimator for the 2×4 table. Let the focus parameter β denote the log odds ratio for interaction and θ denote the log odds ratio between G and E in controls. Then, for a fixed $\theta, \hat{\beta}_{ML}(\theta) \equiv \hat{\beta}_{CO} - \theta$ where $\hat{\beta}_{CO}$ denotes the log odds ratio between G and E in cases. So $\hat{\Delta} = \partial\hat{\beta}_{ML}^\top(\theta)/\partial\theta = -1$. In this special case, the prior covariance $\boldsymbol{A}$ is a positive scalar τ^2, consistent with our previous notation in Section 2. Thus, following (11), the EB-type estimator of β using our general profile likelihood based framework is given by

$$\hat{\beta}_{EB} = \frac{\hat{\sigma}^2_{\hat{\beta}_{ML}}}{\left(\hat{\tau}^2 + \hat{\sigma}^2_{\hat{\beta}_{ML}}\right)}\hat{\beta}^0_{ML} + \frac{\hat{\tau}^2}{\left(\hat{\tau}^2 + \hat{\sigma}^2_{\hat{\beta}_{ML}}\right)}\hat{\beta}_{ML} = \frac{\hat{\sigma}^2_{CC}}{\left(\hat{\theta}^2_{GE} + \hat{\sigma}^2_{CC}\right)}\hat{\beta}_{CO} + \frac{\hat{\theta}^2_{GE}}{\left(\hat{\theta}^2_{GE} + \hat{\sigma}^2_{CC}\right)}\hat{\beta}_{CC},$$

which is exactly what we have proposed in Section 2.

Variance–covariance matrix of the EB estimator: The variance of the proposed estimator in (9) can be obtained by viewing $\hat{\beta}_{EB}$ as a function of the ML estimates, $(\hat{\beta}_{ML}, \hat{\theta}_{ML}, \hat{\beta}^0_{ML})$. The joint asymptotic multivariate normal distribution for these three estimates can be obtained in terms of the associated score functions and information matrices following classical ML theory. An application of the multivariate Taylor's expansion provides the variance–covariance expression for $\hat{\beta}_{EB}$. The derivation and expression of the variance–covariance matrix is deferred to Web Appendix B. The small sample performance of the variance estimator in the simulation setting of Section 3.1 is shown in Web Table 3.

3.1 *Simulation Study with Bivariate Environmental Exposure*

In this section, we design a simulation study involving a binary genetic factor G and two binary environmental exposures E_1 and E_2. The joint distribution of (G, E_1, E_2) among the controls is specified as follows. We assume $P(G = 1) = P(E_1 = 1) = P(E_2 = 1) = 0.3$, and allow E_1 and E_2 to be associated with $OR(E_1, E_2) = 2.0$. We assume G and E_1 are independent with $OR(G, E_1) = 1$, but G and E_2 are associated with $OR(G, E_2) = 1.5$. With the parameters fixed at these values, one can solve a system of equations to obtain the multinomial probability vector corresponding to the eight possible configurations of (G, E_1, E_2). We assume a disease risk model with no main effects for G, E_1, or E_2, but allow for interactions for both E_1 and E_2 with G, with the corresponding log odds ratio parameters being $\beta_{G*E_1} = \beta_{G*E_2} = \log(2)$. Given the control probabilities and the restrictions on the parameters in the disease risk model, one can determine the probabilities for each (G, E_1, E_2) configuration in the case population. We also considered other simulation settings where the disease risk model included main effects (results not shown), the basic pattern of results remain fairly similar.

Table 3 exhibits that the EB method weighs more toward the constrained MLE for estimation of interaction involving G and E_1 for which the independence assumption does in fact hold, but weighs toward the unconstrained MLE for estimation of the interaction involving G and E_2 for which the independence assumption is violated. As a result, it maintains much smaller MSE than the unconstrained MLE for estimation of $G*E_1$ interaction by reducing its variance. It also maintains much smaller MSE than the constrained MLE for estimation of $G*E_2$ interaction by reducing its bias. If one considers the sum of the MSEs corresponding to the two interaction parameters as a performance criterion, the EB estimate has leverage over all the other contenders.

This simulation brings out a major appealing feature of the EB-type estimator. It is often the case that one is considering multiple interaction parameters where the independence assumption may hold for some, but not hold for others, or may be quite ambiguous for a subset. In such situations, one can tacitly avoid specifying which of the independence models are likely to hold and simply use the EB estimator as a data-adaptive solution to the vexing problem of model specification. Remarkably, one can still maintain attractive MSE properties in finite samples without relying on unverifiable model assumptions.

4. Data Analysis

In this section, we apply the proposed methodology to two real datasets, reflecting different degrees of certainty regarding the *G-E* independence assumption. Both examples illustrate the adaptability of the EB estimator depending upon the nature of the *G-E* association present in the data.

4.1 *Analysis of Israeli Ovarian Cancer Data*

This example involves a population-based case–control study of ovarian cancer conducted in Israel, data from which was first reported in Modan et al. (2001) and was then reanalyzed by Chatterjee and Carroll (2005). The main goal of the study

Table 3

Simulation results showing MSE and bias (in parentheses) for estimation of interaction parameters of one genetic factor (G) with two environmental exposures (E_1, E_2). The joint distribution of (G, E_1, E_2) in the controls was specified by the following restrictions: $P(E_1 = 1) = P(E_2 = 1) = P(G = 1) = 0.3, OR_{E_1\,E_2} = 2.0$, $OR_{GE_1} = 1, OR_{GE_2} = 1.5$. The parameters for the disease risk model were set at $\beta_G = \beta_{E_1} = \beta_{E_2} = 0$, and, $\beta_{G*E_1} = \beta_{G*E_2} = \log(2)$. Results are based on 1000 simulated datasets.

		*MSE*1 ($G*E_1$)	*MSE*2 ($G*E_2$)	*MSE*1 + *MSE*2
$n_0 = n_1 = 100$	Dependence	0.46 (0.04)	0.48 (0.10)	0.94
	Independence	0.20 (0.04)	0.39 (0.44)	0.59
	EB	0.29 (0.03)	0.36 (0.24)	0.65
$n_0 = n_1 = 200$	Dependence	0.21 (0.05)	0.21 (0.01)	0.42
	Independence	0.10 (0.02)	0.26 (0.41)	0.36
	EB	0.15 (0.03)	0.16 (0.14)	0.31
$n_0 = n_1 = 500$	Dependence	0.08 (0.01)	0.09 (0.01)	0.17
	Independence	0.04 (0.00)	0.21 (0.41)	0.25
	EB	0.06 (0.00)	0.09 (0.12)	0.15

Table 4

Analysis of Israeli ovarian cancer data: Estimates of the log odds ratio parameters corresponding to each effect is provided, accompanied with 95% confidence intervals[a]

	$\hat{\beta}_{BRCA1/2}$	$\hat{\beta}_{OC}$	$\hat{\beta}_{\text{parity}}$	$\hat{\beta}_{BRCA1/2*OC}$	$\hat{\beta}_{BRCA1/2*\text{parity}}$
Dependence	3.442	−0.051	−0.060	0.049	−0.131
CI	(2.476,4.408)	(−0.108,0.006)	(−0.126,0.006)	(−0.104,0.203)	(−0.373,0.111)
Independence	3.154	−0.051	−0.061	0.086	−0.036
CI	(2.509,3.799)	(−0.102,−0.001)	(−0.125,0.002)	(0.021,0.15)	(−0.141,0.068)
EB	3.270	−0.051	−0.061	0.071	−0.075
CI	(2.40, 4.133)	(−0.108,0.006)	(−0.127,0.005)	(−0.038,0.181)	(−0.282,0.143)

[a]The analysis is adjusted for effects of age, ethnicity, PHB, FHBO, and history of gynecological surgery.

was to examine how mutations in the two major susceptibility genes BRCA1 and BRCA2 may interact with known reproductive risk factors for ovarian cancer, such as number of years of oral contraceptive (OC) use and number of children (parity). Both Modan et al. (2001) and Chatterjee and Carroll (2005) analyzed data from this study assuming independence of BRCA1/2 mutations and the reproductive risk factors in the general population. We revisited the study to explore how the estimates of regression parameters from the previous analyses may change if a certain amount of uncertainty regarding the gene-environment independence assumption was allowed using the proposed EB framework.

Our analysis included 1579 observations in the dataset with 832 cases and 747 controls who did not have bilateral oophorectomy. Similar to Chatterjee and Carroll (2005), we considered fitting a logistic regression model that included main effects for BRCA1/2 mutations (presence/absence), OC, parity, and the interaction terms OC∗BRCA1/2 and parity∗BRCA1/2. The model was adjusted for a set of covariates S that included age (categorized into five groups, by decades), ethnicity (Ashkenazi or non-Ashkenazi), presence of personal history of breast cancer (PHB), family history of breast or ovarian cancer (FHBO, coded as 0 for no history in family, 1 for one breast cancer case in the family, and 2 for one ovarian cancer or two or more breast cancers in family), and history of gynecological surgery. The model for BRCA1/2 mutation frequency is parameterized as

$$\begin{aligned}\text{logit}\{pr(G = 1 \mid E, \boldsymbol{S})\} &= \eta_0 + \eta_{\text{Age}} I(\text{Age} \geq 50) \\ &+ \eta_{Eth} I(\text{Non-Ashkenazi}) + \eta_{PH} I(\text{PHB} = 1) \\ &+ \eta_{1FH} I(\text{FHBO} = 1) + \eta_{2FH} I(\text{FHBO} = 2) \\ &+ \theta_{OC}\text{OC} + \theta_{par}\text{Parity}.\end{aligned}$$

Chatterjee and Carroll (2005) assumed the constrained model $\theta_{OC} = \theta_{\text{par}} = 0$, which implies conditional independence of reproductive risk factors and BRCA1/2 mutation given the stratification factors $\boldsymbol{S}$. Table 4 shows the estimates and 95% confidence intervals for disease log odds ratio parameters of interest under the independence model, dependence model, and using the proposed EB estimator. Under the dependence model, the *G-E* association parameters were estimated as $\hat{\theta}_{OC} = 0.036$ and $\hat{\theta}_{\text{par}}$=0.094. We can notice from Table 4 that the EB point estimate regarding BRCA1/2 ∗ OC interaction is closer to the independence model, whereas EB point estimate regarding the BRCA1/2 ∗ parity interaction is intermediate between the unconstrained and constrained model. The confidence intervals based on the constrained MLEs and EB estimator are noticeably narrower when compared to those obtained from the dependence model.

4.2 *Analysis of Colorectal Adenoma Data*

The second example involves a case–control study of colorectal adenoma, a precursor of colorectal cancer, smoking and

Table 5
Analysis of colorectal adenoma data: Estimates of the log odds ratio parameters corresponding to each effect are provided, accompanied with 95% *confidence intervals*[a]

	$\hat{\beta}_{NAT2=2}$	$\hat{\beta}_{SMK=1}$	$\hat{\beta}_{SMK=2}$	$\hat{\beta}_{NAT2=2*SMK=1}$	$\hat{\beta}_{NAT2=2*SMK=2}$
Dependence	0.833	0.196	1.03	−1.103	−2.784
CI	(0.035,1.632)	(−0.073,0.464)	(0.692,1.367)	(−2.227,0.021)	(−4.569,−0.998)
Independence	0.596	0.176	0.999	−0.766	−2.308
CI	(−0.046,1.239)	(−0.089,0.443)	(0.667,1.332)	(−1.630,0.098)	(−3.885,−0.732)
EB	0.698	0.183	1.009	−0.923	−2.531
CI	(−0.031,1.426)	(−0.083, 0.449)	(0.678,1.339)	(−1.956, 0.111)	(−4.214,−0.848)

[a]The analysis is adjusted for effects of age, gender, and family history of colorectal cancer.

NAT2, a gene that is believed to play an important role in metabolism of smoking-related carcinogens. In this study, a total of 772 left-sided prevalent advanced adenoma cases and 777 gender and ethnicity-matched controls were selected from the screening arm of the large ongoing Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial at the National Cancer Institute, United States (Gohagan et al., 2000; Hayes et al., 2005). Subjects selected in the case–control study were genotyped for six single nucleotide polymorphisms that have been related to *NAT2*-acetylation activity in previous laboratory studies. Based on the genotypes, subjects were assigned an acetylation phenotype as "slow" ($NAT2 = 0$), "intermediate" ($NAT2 = 1$), or "rapid" ($NAT2 = 2$). Baseline questionnaire data were used to categorize subjects as "never" ($SMK = 0$), "former" ($SMK = 1$), or "current" ($SMK = 2$) smokers. Results from standard logistic regression analysis of this data has been recently reported by Moslehi et al. (2006). We considered reanalysis of this study in the proposed EB framework. We restricted the analysis to Caucasian subjects who had complete NAT2-phenotype information, resulting in a total of 610 cases and 605 controls. We considered fitting a logistic regression model with main effects of smoking, NAT2 (categorized as rapid or not), and their interactions. The model was adjusted for co-factors $\boldsymbol{S}$ that included age, gender, and family history of colorectal cancer ($FHCO = 1$ for yes, 0 for no). The prevalence of NAT2 rapid acetylation phenotype was modeled as

$$\begin{aligned}\text{logit}\{P(NAT2 = 2|E, \boldsymbol{S})\} &= \eta_0 + \eta_{FHCO} I(FHCO) \\ &+ \eta_{\text{gender}} I(Male) + \theta_{SMK1} I(SMK = 1) \\ &+ \theta_{SMK2} I(SMK = 2).\end{aligned}$$

In this dataset, there seems to be much less certainty about the independence of NAT2 and smoking, with $\theta_{SMK1} = 0.340$; $\theta_{SMK2} = 0.495$. Results in Table 5 show estimates of the interaction between NAT2 rapid enzymatic phenotype and current smokers ($NAT2 = 2 * SMK = 2$) is highly significant under all models, whereas the interaction between NAT2 rapid enzymatic phenotypes with former smokers ($NAT2 = 2 * SMK = 1$) is not significant under any model. The EB estimates of interaction parameters for this dataset are not quite close to the ones obtained from the independence model. The EB confidence intervals are considerably narrower compared to the corresponding intervals from the dependence model, reflecting the combined efficiency-robustness feature of the EB estimator.

5. Discussion

EB (Efron and Morris, 1972; Morris, 1983; Efron, 1993; Carlin and Louis, 2000) is a pragmatic Bayesian paradigm, steering between the extreme Bayesian and frequentist standpoints. In the context of the problem of relaxing gene-environment independence assumption, the proposed EB-type approach has a natural appeal and interpretation, powered with an extremely straightforward ML-based computation. This makes the method readily available and implementable to the practitioner. We believe, for example, the simple closed form expression for the estimate of interaction between a binary genetic and a binary environmental exposure should facilitate the use of the method for very large-scale studies such as a genome-wide scan. We also observe that although the estimator is conceived from a Bayesian standpoint, it is simply a functional of the observed data and can thus be viewed as a novel frequentist estimator. Our simulation studies involving fixed parameter settings indicate that the estimator has good frequentist properties in the sense of maintaining low MSEs across different scenarios of gene-environment dependence. The proposed methodology can be easily adapted to construct EB-type shrinkage estimator assuming a nonzero, but known, prior mean for the gene-environment log odds ratio parameters. For commonly studied combinations, such as *NAT2* and *smoking*, information on such prior mean may be gathered by meta-analysis of the gene-environment association from previous studies (see, e.g., Marcus et al., 2000).

As discussed in the introduction, practitioners may find it natural to resolve the bias versus efficiency issue by deciding between the case-only and case–control estimators depending on a statistical test of the independence assumption $\theta_{GE} = 0$ using the control sample. This "two-stage" method essentially leads to a weighted estimator for the interaction parameter with weights being 0–1 random variable indicating the acceptance/rejection of the test of the null hypothesis of independence. Our simulation studies indicate that the discrete weights of the two-stage method generally lead to substantially larger bias and MSEs than those obtained using the EB-weights which depend on θ_{GE} in a continuous fashion. Moreover, obtaining a proper variance estimator for the two-stage estimator, accounting for the uncertainty of the decision rule associated with the hypothesis testing of independence, can be fairly complex. A naive approach that uses the

standard case–control or case-only variance estimator depending on which of the two estimators is being used for a given study leads to underestimation of the variance of the whole procedure. The resulting test of interaction could have highly inflated type I error (Albert et al., 2001).

It is important to note that the proposed estimator, although it performs very well in terms of MSE, can have modest bias in parameter estimates when the gene-environment independence assumption is violated and for modest sample sizes. As a result, the associated confidence intervals can also have less than nominal coverage (see Web Tables 2 and 3). Thus, if bias is used as the primary criterion for evaluating the methods, then standard logistic regression 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 modest, e.g. $\exp(\theta_{GE}) \leq 1.2$ or $\exp(\theta_{GE}) \geq 0.8$, then the bias of the proposed estimator is quite small and coverage of the associated confidence intervals is close to the desired level (see first block of Web Table 3 and Web Table 4). Empirical studies suggest that violation of gene-environment independence, when it occurs, would likely to be modest in most situations (Liu, Fallin, and Kao, 2004). Thus, we believe that the proposed method, overall, is a promising approach for investigation of gene-environment interaction from case–control studies.

Alternative estimation methods are possible within the stochastic framework we introduced for relaxation of the independence assumption. In the simulation studies described in Table 1, we considered an "IL" approach that estimates all of the parameters of the model jointly from the case–control likelihood, after integrating it with respect to the "random effect" parameter θ_{GE}. We found the IL approach does not perform any better than the simpler EB estimator in terms of MSE. The bias of the IL approach, however, can be smaller under violation of gene-environment independence. Given that the IL approach can be computationally quite complex, especially when there are multiple gene-environment dependence parameters, it would be of future research interest to explore whether there is an alternative estimator that would be computationally simple and yet would be able to achieve smaller bias than the EB estimator.

To develop the EB-type estimator for the general logistic regression model, we approximated the prior variance for the function $\beta(\theta)$ assuming a simple linear Taylor's series approximation for this function in θ. In the simple case involving a binary G and binary E, we have shown that $\beta(\theta)$ is exactly linear in θ. Although in general we do not have any such theoretical result about how good the linear approximation may be, our simulation studies involving multiple environmental exposures indicate the proposed estimator with the assumed linear approximation performs well in appropriately "shrinking" the *G-E* interaction estimates toward constrained or unconstrained model depending on the true nature of gene-environment dependence. Further, in principle, one can also improve the accuracy of the approximation by considering a second-order Taylor's series approximation to the function $\beta(\theta)$. We observe that under such approximations, the prior variance of $\beta(\theta)$ given that $\theta \sim MVN(0, \tau)$ can be computed in terms of second-, third-, and fourth-order moments of normal distribution and the first and second derivative of the function $\beta(\theta)$. The remainder of the calculation would remain identical as the proposed EB estimator.

The proposed "profile-likelihood-empirical Bayes" framework has other potential applications for analysis of case–control studies when certain types of covariate distributional constraints are likely, but not certain. The same framework, for example, can be used to exploit the constraint of HWE for genetic association studies. In this context, development of the EB estimator would first require specifying an "unconstrained" model for the genotype distribution in which the "constraint" of HWE would be a special case. The MLEs of genetic odds ratio parameters under the constrained and unconstrained models can be then combined based on the estimate(s) of certain index parameter(s) that would measure the magnitude of departure of the "unconstrained" genotype distribution from HWE.

The proposed framework also raises a number of interesting theoretical issues including how it relates to a proper full Bayes procedure. Intuitively, a noninformative or minimally informative prior on β, after a possible orthogonalization (Tibshirani, 1989) of the parameter space for (β, θ), may lead to approximately similar inference. An in-depth, rigorous examination of this connection is needed in the future.

In conclusion, the proposed methodology provides a promising solution to the bias versus efficiency dilemma faced in case–control studies due to the assumption of gene-environment independence assumption. Further, the general framework we provide could be useful for resolving similar issues in other areas of epidemiologic studies.

6. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2, 3, and 5 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

This research was supported by the Intramural Program of the National Institute of Health. The research of BM was also partially supported by NSA young investigator grant H98230-06-1-0033. Computer codes using the `MATLAB` software for implementing the analysis is available at `http://dceg.cancer.gov/about/staff-bios/chatterjee-nilanjan#software`. The authors would like to thank Drs Sholom Wacholder, Mike Boenhke, and one anonymous reviewer for valuable comments leading to substantial improvement of the manuscript.

References

Albert, P. S., Ratnasinghe, D., Tangrea, J., and Wacholder, S. (2001). Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology* **154,** 687–693.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B* **32,** 283–301.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer Verlag.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*, 2nd edition. Boca Raton, FL: Chapman and Hall/CRC Press.

Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* **92,** 399–418.

Chen, J. and Chatterjee, N. (2007). Exploiting Hardy-Weinberg equilibrium for efficient screening of single SNP associations from case-control studies. *Human Heredity* **63,** 196–204.

Cox, D. R. (1975). A note on partially Bayes inference and the linear model. *Biometrika* **62,** 651–654.

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80,** 3–26.

Efron, B. and Morris, C. (1972). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika* **59,** 335–347.

Epstein, M. P. and Satten, G. A. (2003). Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics* **73,** 1316–1329.

Gohagan J. K., Prorok P. C., Hayes R. B., and Kramer B. S. (2000). Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Project Team. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Controlled Clinical Trials* **21(6 suppl.),** 273S–309S.

Greenland, S. (1993). Methods for epidemiologic analyses of multiple exposures: A review and comparative study of maximum likelihood, preliminary-testing, and empirical Bayes regression. *Statistics in Medicine* **12,** 717–736.

Hayes R. B., Sigurdson A., Moore L., Peters U., Huang W. Y., Pinsky P., Reding D., Gelmann E. P., Rothman N., Pfeiffer R. M., Hoover R. N., and Berg C. D. (2005). Methods for etiologic and early marker investigations in the PLCO Trial. *Mutation Research* **592,** 147–154.

Lin, D. Y. and Zeng, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the American Statistical Association* **101,** 89–104.

Liu, X., Fallin, M. D., and Kao, W. H. (2004). Genetic dissection methods: Designs used for tests of gene-environment interaction. *Current Opinions in Genetics and Development* **14,** 241–245.

Marcus, P. M., Hayes, R. B., Vineis, P., et al. (2000). Cigarette smoking: N-acteyltransferease 2 acetylation status, and bladder cancer risk: A case series meta-analysis of a gene-environment interaction. *Cancer Epidemiology, Biomarkers and Prevention* **9,** 461–467.

Modan, B., Hartge, P., Hirsh-Yechezkel, G., et al. (2001). Parity, oral contraceptives and the risk of ovarian cancer among carriers and non-carriers of a BRCA1 or BRCA2 mutation. *New England Journal of Medicine* **345,** 235–240.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association* **78,** 47–55.

Moslehi, R., Chatterjee, N., Church, T. R., Chen, J., Yeager, M., Weissfield, J., Hein, D. W., and Hayes, R. B. (2006). Cigarette smoking n-acetyltransferase genes and the risk of advanced colorectal adenoma. *Pharmacogenomics* **7,** 819–829.

Piegorsch, W. W., Weinberg, C. R., and Taylor, J. (1994). Nonhierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine* **13,** 153–162.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66,** 403–411.

Satten, G. A. and Epstein, M. P. (2004). Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genetic Epidemiology* **27,** 192–201.

Satten, G. A. and Kupper, L. L. (1993). Inferences about exposure-disease associations using probability-of-exposure information. *Journal of the American Statistical Association* **88,** 200–208.

Schmidt, S. and Schaid, D. J. (1999). Potential misinterpretation of the case-only study to assess gene-environment interaction. *American Journal of Epidemiology* **150,** 878–885.

Spinka, C., Carroll, R. J., and Chatterjee, N. (2005). Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genetic Epidemiology* **29,** 108–127.

Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika* **76,** 604–608.

Umbach, D. M. and Weinberg, C. R. (1997). Designing and analysing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine* **16,** 1731–1743.

Received September 2006. *Revised September* 2007.
Accepted October 2007.