University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2006 *Paper* 65

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

Bhramar Mukherjee*

*University of Michigan, bhramar@umich.edu

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

Bhramar Mukherjee

Abstract

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, modern "retrospective" methods, including the celebrated "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 approach to analyze case-control data that can relax the gene-environment independence assumption using an empirical Bayes (EB) framework. In the special case, involving a binary gene and a binary exposure, the framework leads to an estimator of the odds-ratio interaction 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 EB estimator and its variances within the retrospective maximum-likelihood framework developed by Chatterjee and Carroll (2005). We conduct simulation studies to investigate the mean-squared-error of the proposed estimator in both fixed and random parameter settings. We also illustrate the application of this methodology using two real data examples. Both simulated and real data examples suggest that the proposed estimator strikes an excellent balance between bias and efficiency depending on the true nature of the gene-environment association and the sample size for a given study.

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

BHRAMAR MUKHERJEE[1] AND NILANJAN CHATTERJEE[2]

[1]Department of Biostatistics, University of Michigan
Ann Arbor, MI 48109
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
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, modern "retrospective" methods, including the celebrated "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 approach to analyze case-control data that can relax the gene-environment independence assumption using an empirical Bayes (EB) framework. In the special case, involving a binary gene and a binary exposure, the framework leads to an estimator of the odds-ratio interaction 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 EB estimator and its variances within the retrospective maximum-likelihood framework developed by Chatterjee and Carroll (2005). We conduct simulation studies to investigate the mean-squared-error of the proposed estimator in both fixed and random parameter settings. We also illustrate the application of this methodology using two real data examples. Both simulated and real data examples suggest that the proposed estimator strikes an excellent 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; Bayes; 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 et al. (1994) noticed that under G-E independence and assuming 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 estimate 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 analysis of case-control data under the gene-environment independence assumption in a general setting that may involve continuous exposures, non-rare 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. 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 (Epstein and Satten, 2003; Satten and Epstein 2004; Spinka et al, 2005; Lin and Zeng, 2006).

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 Hardy-Weinberg-Equilibrium are violated (Albert

et al, 2001; Chatterjee and Carroll, 2005; Satten and Epstein 2004; 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 type 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 use the outcome of that test to decide whether to the 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 vs 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. The estimate of this uncertainty parameter is then used, in an "empirical Bayes (EB)" way, to combine the two sets of estimates of odds-ratio parameters, one obtained assuming G-E independence and the other obtained allowing a general model for G-E dependence.

In Section 2, we consider a simple scenario involving a binary G and a binary E, where the EB 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 sam-

ples the proposed estimator strikes an excellent 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 composite estimators for all of the parameters of a general logistic regression model using the retrospective maximum-likelihood framework developed by Chatterjee and Carroll (2005). Further simulation studies are conducted to study 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 compelling evidence of how the EB estimate is tracking the maximum likelihood estimates 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.

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 (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 1.

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 case and control 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, $\boldsymbol{r}_0 \sim$ Multinomial$(n_0, \boldsymbol{p}_0)$ and $\boldsymbol{r}_1 \sim$ Multinomial$(n_1, \boldsymbol{p}_1)$. 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,

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

The assumption of G-E independence, together with the rare disease approximation implies $\theta_{GE} = 0$. 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, $\theta_{GE} \sim N(0, \tau^2)$, 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 $\hat{\theta}_{GE} = \log \{(r_{000}r_{011})/(r_{001}r_{010})\}$. Standard likelihood theory implies that, given θ_{GE}, $\hat{\theta}_{GE} \sim N(\theta_{GE}, \sigma^2_{\hat{\theta}_{GE}})$, where an estimate of the asymptotic variance is given by $\hat{\sigma}^2_{\hat{\theta}_{GE}} = \sum_{g=0}^{1}\sum_{e=0}^{1}(1/r_{0ge})$. Unconditioning on θ_{GE}, it follows that marginally $\hat{\theta}_{GE} \sim N(0, \tau^2 + \sigma^2_{\hat{\theta}_{GE}})$. Thus, based on the marginal likelihood 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 propose to use 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 $\beta = \log(\psi)$.

With this Bayesian framework in mind, we now propose a composite 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

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

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). Employing standard asymptotic theory, the asymptotic variances of these two estimators can be obtained as $\hat{\sigma}^2_{CC} = \sum_{d=0}^{1}\sum_{g=0}^{1}\sum_{e=0}^{1}(1/r_{dge})$ and $\hat{\sigma}^2_{CO} = \sum_{g=0}^{1}\sum_{e=0}^{1}(1/r_{1ge})$. Consider the following weighted estimator of the interaction parameter:

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

The form of the estimator is motivated by the expression for the posterior mean obtained in a conjugate analysis under a normal-normal model (Berger, 1985, p131), with the prior variance substituted by an estimate obtained from the marginal likelihood in the spirit of Morris (1983). 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 partially Bayes inference introduced by Cox (1975).

We observe that as $\hat{\tau}^2 = \hat{\theta}^2_{GE} \to 0$, i.e. as the data provides 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, $\hat{\beta}_{EB} \to \hat{\beta}_{CC}$. Since $\hat{\beta}_{CC} = \hat{\beta}_{CO} - \hat{\theta}_{GE}$, 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 EB estimator under varying scenarios of G-E association.

2.1 Simulation study for the 2×4 table

Though the estimate $\hat{\beta}_{EB}$ is postulated in a Bayesian framework, it is purely a functional of the data (namely, the multinomial counts, $\boldsymbol{r}_0$ and $\boldsymbol{r}_1$). 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 frequentist paradigm. We then proceed to study the performance of this estimator in a random parameter setting, motivated by a real scenario where one could effectively elicit a plausible value for the prior hyperparameter τ based on published data.

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 could obtain the control probability vector $\boldsymbol{p}_0$ by solving a system of equations. We then set the values of OR_{10}, OR_{01} and ψ, which together with $\boldsymbol{p}_0$, defines 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 EB 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.

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 10000 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 sample. The EB estimator also has a clear edge over the two-stage estimator in terms of bias and MSE, especially in small samples.

In the second set of simulation, we attempt to generate a stochastic model for uncertainty regarding the G-E independence assumption following Marcus *et al.* (2000), who published data on association between N-Acetyltranferase 2 (NAT2) acetylation status (slow vs. rapid), and smoking (ever vs. never) using the controls from 11 different case-control studies conducted in US, Europe, India and Japan (Table 3, Figure 1). Study to study variation in the odds ratios between NAT2 and smoking was noted though none of the associations were found to be statistically significant. It is evident that though the overall G-E independence assumption holds (mean over the 11 sites=0.07), there is variation across study sites (sd=0.35). Treating this as a replica of the distribution for θ_{GE}, in each simulated dataset, we generate θ_{GE} at random from a normal distribution with mean 0.07 and standard deviation 0.35, while values of all other simulation parameters (P_G,P_E, OR_{10}, OR_{01} and OR_{11}) are held fixed.

Results shown in Table 4 bring out certain interesting features of the three estimators. Even though the overall G-E independence is reasonable, with a variation about the mean, the case-only estimator becomes inferior, whereas our proposed weighted estimator continues to adapt itself to model this uncertainty while maintaining significant efficiency gain. In Table 4, we also included a modification of the EB estimator, namely EB-TRUE, where instead of estimating the prior variance τ^2 by its marginal MLE, we substituted τ^2 by its true value 0.35. The results illustrate that the proposed EB estimator is marginally inferior to EB-TRUE, for $n_0 = n_1 = 100$, but performs nearly as well as EB-TRUE for moderately larger sample sizes.

Variance of the proposed estimator: In the following, we propose a method to obtain an asymptotic variance expression for $\hat{\beta}_{EB}$. Since $\hat{\sigma}^2_{cc} \rightarrow 0$ 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 $\sqrt{n}$-asymptotic approximation of the EB 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 the delta theorem 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,

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

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 supplementary Table 1).

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 factor(s) such as ethnicity which could be related to both G and E. They consider the following factorization of the retrospective likelihood,

$$L^R = \text{pr}(G, E|D) = \frac{\text{pr}(D|G, E, \boldsymbol{S})\text{pr}(G|E, \boldsymbol{S})\text{pr}(E, \boldsymbol{S})}{\sum_{\text{G,E},\boldsymbol{s}} \text{pr}(D|G, E, \boldsymbol{S})\text{pr}(G|E, \boldsymbol{S})\text{pr}(E, \boldsymbol{S})}. \tag{5}$$

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 $\text{pr}(D = 1|G, E, \boldsymbol{S}) = H\{\gamma_0 + m(G, E, \boldsymbol{S}; \gamma_1)\}$ where $H(u) = (1 + \exp(-u))^{-1}$ and $m(\cdot)$ a known but arbitrary function. The joint distribution function for $(E, \boldsymbol{S})$ is allowed to remain completely unrestricted (non-parametric). Under G-E independence, conditional on $\boldsymbol{S}$, $\text{pr}(G|E, \boldsymbol{S}) = \text{pr}(G|\boldsymbol{S})$. Assuming a binary genetic factor G, consider a

logistic model of the form

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

Extension to a general categorical G via a multinomial logistic link is immediate. We will refer to (6) as the independence model, or the constrained model. Without the assumption of G-E independence, one can expand the model in (6) to

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

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 maximum-likelihood estimates for the parameters $\omega = (\boldsymbol{\gamma}, \boldsymbol{\eta})$ under model (6) as well as those for $\omega = (\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\theta})$ 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 $Z = (E, S)$ to remain completely non-parametric, can be obtained by maximizing a simpler pseudo-likelihood of the form $L^* = \text{pr}(D, G|E, S, R = 1)$, 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 $\boldsymbol{\beta}$=($\boldsymbol{\gamma}$, $\boldsymbol{\eta}$) under the unconstrained and constrained models will be denoted by $\hat{\boldsymbol{\beta}}_{ML}$ and $\hat{\boldsymbol{\beta}}^0_{ML}$, respectively.

Before we proceed to form the EB 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 $\text{MVN}_m(0, \boldsymbol{\tau})$, a m-dimensional zero-mean multivariate normal distribution with variance-covariance matrix $\boldsymbol{\tau}$. 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 maximum-likelihood estimate 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 $\phi = f(\theta)$, when θ has a prior $N(0, \boldsymbol{\tau})$. By applying the delta theorem, the prior on ϕ could be approximated as $\phi \sim N(f(0), f'(0)\boldsymbol{\tau}\{f'(0)\}^\top)$, where $f'(\theta)$ denotes the gradient matrix of f with respect to θ. 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 $\boldsymbol{\tau}$ is given by

$$\hat{\phi} = f'(0)\boldsymbol{\tau}\{f'(0)\}^\top \left[\hat{V}_\phi + \{f'(0)\}\boldsymbol{\tau}\{f'(0)\}^\top\right]^{-1} f(\hat{\theta}_{ML}) + \hat{V}_\phi \left[\hat{V}_\phi + f'(0)\boldsymbol{\tau}\{f'(0)\}^\top\right]^{-1} f(0). \tag{8}$$

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 $\boldsymbol{\tau}$ as,

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

where $\Delta = \partial\beta^\top(\theta)/\partial\theta$, is the gradient matrix of dimension $m \times p$ evaluated at $\theta = 0$. Note that $\Delta^\top\boldsymbol{\tau}\Delta$ 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

$$\hat{\Delta} = \frac{\partial \hat{\beta}_{ML}^T}{\partial \theta}(\theta = 0) = -\boldsymbol{I}_{\theta\beta}(\theta = 0)\left\{\boldsymbol{I}_{\beta\beta}(0)\right\}^{-1}. \tag{10}$$

Here $\boldsymbol{I}_{\theta\beta}$ and $\boldsymbol{I}_{\beta\beta}$ denote suitable information matrices under the unconstrained model. In alignment with the empirical Bayes spirit, we now estimate the prior hyperparameter $\boldsymbol{\tau}$ 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 EB estimate is given by

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

Computationally, this requires only fitting the constrained model and the unconstrained model and extracting the variance covariance components for the unconstrained model and evaluating it at $\theta = 0$.

Revisiting the simple estimator in the 2×4 case: Now consider our proposed estimator in the 2×4 table. Let the focus parameter β denote the log odds ratio for interaction and θ denotes the log odds ratio between G and E in controls. Then, for a fixed θ, $\hat{\beta}_{ML}(\theta) \equiv \hat{\beta}_{CO} - \theta$ where $\hat{\beta}_{CO}$ denotes the log odds ratio between G and E in cases. So $\hat{\Delta} = \partial\hat{\beta}_{ML}(\theta)/\partial\theta = -1$.

Thus, following (11), the "profile likelihood-Empirical Bayes" estimate of $\boldsymbol{\beta}$ using our general framework is given by

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

which is exactly what we have proposed in Section 2.

Variance-Covariance matrix of the EB estimate: The variance of the proposed estimator in (9) can be obtained by viewing $\hat{\beta}_{EB}$ as a function of the ML estimators, $(\hat{\boldsymbol{\beta}}_{ML}, \hat{\theta}_{ML}, \hat{\boldsymbol{\beta}}^0_{ML})$. 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 delta theorem provides the variance-covariance expression for $\hat{\beta}_{EB}$. The derivation and expression of the covariance matrix is deferred to the appendix. The small sample performance of the variance estimator in the simulation setting of Section 3.1 is shown in supplementary Table 2.

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 following. 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.

The simulation results in Table 5 exhibit that the EB estimate is closer to the constrained MLE in determining $G * E_1$ interaction, where independence assumption does in fact hold, whereas in estimating $G * E_2$ interaction, for which the independence assumption is violated, it is closer to the unconstrained MLE. For both the interaction parameters, EB has smaller MSE compared to the unconstrained ML estimator. The constrained ML estimator assuming G-E independence performs very poorly for estimating $G * E_2$ interaction. If one considers the sum of the MSE's 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 estimate. 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 the data examples present strong evidence for 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

The first 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 re-analyzed by Chatterjee and Carroll (2005). The main goal of the study 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 re-visited the study to explore how the estimates of regression parameters from the previous analyses may change if 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 $\boldsymbol{S}$ that included age (categorized into 5 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 gynaecological surgery. The model for BRCA1/2 mutation frequency is parameterized as

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

Chatterjee and Carroll (2005) assumed the constrained model $\theta_{OC} = \theta_{par} = 0$, which implies conditional independence of reproductive risk factors and BRCA1/2 mutation given the stratification factors $\boldsymbol{S}$. Table 6 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 $\hat{\theta}_{par}$=0.094. We can notice from Table 6 that EB inference regarding BRCA1/2$*$OC interaction is closer to the independence model, whereas EB inference regarding the BRCA1/2$*$parity interaction is intermediate between the unconstrained and constrained model. This could be expected of the EB estimate as the independence assumption is less certain for BRCA1/2 and parity based on the current data. The independence model and the EB estimate produce significant BRCA1/2*OC interaction estimate, while the unconstrained model fails to detect significance. On the other hand, the BRCA1/2*parity interaction is not significant under any model, though 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 *NAT2*, a genomic region 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, USA (Gohagan et al., 2000; Hayes et al., 2005). Subjects selected in the case-control study were genotyped for six single neucleotide polymorphisms (SNPs) 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 re-analysis of this study in the proposed EB framework. We restricted the analysis to Caucasian subjects who has 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 modelled as,

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

In this dataset, there seems to be much less certainty about the independence of NAT2 and smoking, with $\theta_{NAT2=2,SMK=1}$=0.340; $\theta_{NAT2=2,SMK=2}$=0.495, (CI=(-0.425,1.415)); Notice that these values can be viewed as typical realizations from the distribution elicited via the Marcus *et al.* (2000) data (Figure 1). 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

Empirical Bayes (Efron and Morris, 1972; Morris, 1983; Efron, 1993; Carlin and Louis, 2000) is a pragmatic Bayesian paradigm, steering between the extreme Bayesian and frequentist standpoint. In the context of the problem of relaxing gene-environment independence assumption, the proposed Empirical Bayes (EB) approach has a natural appeal and interpretation, powered with an extremely straightforward maximum likelihood 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 would facilitate the use of the method for very large-scale studies such as a genomewide scan. We also observe that though 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 setting indicate that the estimator has excellent frequentist properties in the sense of maintaining low mean-squared-errors across different scenarios of gene-environment dependence.

Our simulation study also reveals some interesting features of the case-only estimator. When we simulated case-control studies from a fixed population (Table 2), for which the gene-environment independence assumption holds, the case-only estimate of interaction had the smallest mean-squared errors among all of the methods considered. However, when we simulated case-control studies from different populations allowing for some study-to-study variability in the gene-environment

distribution (Table 3), the case-only estimator became much inferior to the case-control estimator even though the independence assumption was satisfied in the overall super-population. In contrast, the proposed EB estimator always maintained significantly smaller or similar MSE as the case-control estimator. Given that this kind of study-to-study variability may be quite natural in practice, as seen in the data published by Marcus et al. (2000), the performance of the EB estimator in the random parameter setting seems very promising.

As discussed in the introduction, practitioners may find it natural to resolve the bias vs 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 an 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 two-stage method generally leads to substantially larger bias and mean-squared-errors 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 estimates 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).

Another simple way to combine the case-only and case-control estimator would be to use a Bayesian Model Averaging (BMA) approach (Madigan and Raftery, 1994) where the weights for the two estimators could be approximated based on the Bayesian Information Criterions (BIC) for the constrained and unconstrained retrospective likelihoods. Our simulation studies indicate (results not included) that the BMA approach has often larger bias and MSE than the EB estimator because the former method often attaches more weight to the constrained estimator due to the large penalty given in the BIC criterion for each extra parameter.

The proposed "empirical-Bayes-profile-likelihood" framework has other potential applications for analysis of case-control studies when certain type of covariate distributional constraints are likely, but not certain. The same framework, for example, can be used to exploit the constraint of Hardy-Weinberg-Equilibrium 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 maximum-likelihood estimates 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 non-informative 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 future.

In conclusion, the proposed methodology provides a promising solution to the bias vs 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.

ACKNOWLEDGEMENTS: This research was supported by the Intramural program of the National Institute of Health. The research of Bhramar Mukherjee 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/people/ChatterjeeNilanjan.html#software.

References

Albert P. S., Ratnasinghe D., Tangrea J., Wacholder S. (2001). Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology* **154**, 687–93.

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 date analysis.*, 2nd edn. Chapman and Hall/ CRC Press, Boca Raton, FL.

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

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.

Efron, B., and Morris, C. (1972), Limiting the risk of Bayes and empirical Bayes estimators – Part II: The empirical Bayes case, *Journal of the American Statistical Association* **67**, 130–139.

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., 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., 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.

Madigan D. M. and Raftery A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's Window. *Journal of the American Statistical Association* **89**, 1335–1346.

Marcus *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 *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., Hayes, R. B. (2006). Cigarette smoking n-acetyltransferase genes and the risk of advanced colorectal adenoma. *Pharmacogenomics* (In Press).

Piegorsch, W. W., Weinberg, C. R. and Taylor, J. (1994). Non hierarchical 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.

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.

Appendix

Variance approximation for the EB estimate in multivariate setting

We first derive the joint asymptotic distribution of the MLE's $\kappa_{ML} = (\hat{\beta}_{ML}, \hat{\theta}_{ML}, \hat{\beta}^0_{ML})^\top$ obtained from models (6) and (7). Let $\boldsymbol{I}$ ($(p+m)\times(p+m)$) and $\boldsymbol{I}^0$ ($p\times p$) denote the observed information matrices for the unconstrained and the constrained models respectively. Then, asymptotically,

$$\begin{aligned}\begin{pmatrix}\hat{\beta}_{ML}\\ \hat{\theta}_{ML}\end{pmatrix} &= \sqrt{n}\boldsymbol{I}^{-1}\sum_{i=1}^{n} U_{\beta,\theta}(D_i, G_i, E_i, \boldsymbol{S}_i) + o_p(n^{-1/2}) \qquad \text{and,}\\ \hat{\beta}^0_{ML} &= \sqrt{n}(\boldsymbol{I}^0)^{-1}\sum_{i=1}^{n} U^0_{\beta}(D_i, G_i, E_i, \boldsymbol{S}_i) + o_p(n^{-1/2}),\end{aligned}$$

where $U_{\beta,\theta}(D_i, G_i, \boldsymbol{S}_i)$, and $U^0_\beta(D_i, G_i, E_i, \boldsymbol{S}_i)$ denote the individual score functions for subject i corresponding to the unconstrained and the constrained MLE with n denoting the total sample size. Consider the partitioning $\boldsymbol{I}^{-1} = \left(\boldsymbol{W}_1^\top, \boldsymbol{W}_2^\top\right)^\top$ where $\boldsymbol{W}_1$ is a $p\times(p+m)$ matrix and $\boldsymbol{W}_2$ is a $m\times(p+m)$ matrix. Further, let $\boldsymbol{W}^0 = (\boldsymbol{I}^0)^{-1}$, $U_i = U_{\beta,\theta}(D_i, G_i, \boldsymbol{S}_i)$ and $U^0_i = U^0_\beta(D_i, G_i, E_i, \boldsymbol{S}_i)$. The asymptotic variance- covariance matrix of the vector of MLE's κ_{ML} can be represented as

$\Sigma_{\kappa_{ML}} =$

$$\begin{pmatrix} \boldsymbol{W}_1\text{Var}(\sum_{i=1}^n U_i)\boldsymbol{W}_1^\top & \boldsymbol{W}_1\text{Var}(\sum_{i=1}^n U_i)\boldsymbol{W}_2^\top & \boldsymbol{W}_1\text{Cov}(\sum_{i=1}^n U_i, \sum_{i=1}^n U^0_i)(\boldsymbol{W}^0)^\top \\ \boldsymbol{W}_2\text{Var}(\sum_{i=1}^n U_i)\boldsymbol{W}_1^\top & \boldsymbol{W}_2\text{Var}(\sum_{i=1}^n U_i)\boldsymbol{W}_2^\top & \boldsymbol{W}_2\text{Cov}(\sum_{i=1}^n U_i, \sum_{i=1}^n U^0_i)(\boldsymbol{W}^0)\top \\ \boldsymbol{W}^0\text{Cov}(\sum_{i=1}^n U^0_i, \sum_{i=1}^n U_i)(\boldsymbol{W}_1)^\top & \boldsymbol{W}^0\text{Cov}(\sum_{i=1}^n U^0_i, \sum_{i=1}^n U_i)\boldsymbol{W}_2^\top & \boldsymbol{W}^0\text{Var}(\sum_{i=1}^n U^0_i)(\boldsymbol{W}^0)^\top \end{pmatrix}.$$

Assuming that the first n_0 subjects among a total of n selected subjects are controls and remaining $n-n_0$ subjects are cases, the covariances appearing in the above matrix can be computed under the case-control sampling scheme as

$$\text{Cov}\Big(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i\Big) = n_0\Big(\sum_{i=1}^{n_0} X_iY_i^\top/n_0 - \overline{X}_{control}\overline{Y}^\top_{control}\Big) + n_1\Big(\sum_{i=n_0+1}^{n} X_iY_i^\top/n_1 - \overline{X}_{case}\overline{Y}^\top_{case}\Big).$$

where $\overline{X}_{case}$ ($\overline{X}_{control}$) and $\overline{Y}_{case}$ ($\overline{Y}_{control}$) denote the average of X and Y among cases and controls respectively.

Now, we write,

$$\begin{aligned}\hat{\beta}_{EB} &= g\left(\hat{\beta}_{ML}, \hat{\theta}_{ML}, \hat{\beta}^0_{ML}\right) \\ &= \hat{\Delta}^\top\hat{\theta}_{ML}\hat{\theta}^\top_{ML}\hat{\Delta}\left(\hat{V}_{\hat{\beta}_{ML}} + \hat{\Delta}^\top\hat{\theta}_{ML}\hat{\theta}^\top_{ML}\hat{\Delta}\right)^{-1}\hat{\beta}_{ML} + \hat{V}_{\hat{\beta}_{ML}}\left(\hat{V}_{\hat{\beta}_{ML}} + \hat{\Delta}^\top\hat{\theta}_{ML}\hat{\theta}^\top_{ML}\hat{\Delta}\right)^{-1}\hat{\beta}^0_{ML}. \\ &= \hat{\beta}_{ML} - \hat{V}_{\hat{\beta}_{ML}}\left(\hat{V}_{\hat{\beta}_{ML}} + \hat{\Delta}^\top\hat{\theta}_{ML}\hat{\theta}^\top_{ML}\hat{\Delta}\right)^{-1}(\hat{\beta}_{ML} - \hat{\beta}^0_{ML}), \qquad (12)\end{aligned}$$

Then, by the multivariate delta theorem, the approximate variance-covariance matrix of $\hat{\beta}_{EB}$ will be given by

$$\left\{g'\left(\hat{\beta}_{ML}, \hat{\theta}_{ML}, \hat{\beta}^0_{ML}\right)\right\}^\top \Sigma_{\kappa_{ML}}\, g'\left(\hat{\beta}_{ML}, \hat{\theta}_{ML}, \hat{\beta}^0_{ML}\right),$$

where g' is the gradient matrix of g with respect to $(\hat{\beta}_{ML}, \hat{\theta}_{ML}, \hat{\beta}^0_{ML})$. In defining the gradients, we follow the convention that the derivative of the vector function $\boldsymbol{f}(u)$ (of length r, say) with respect to a vector u (of length s, say) denotes a matrix of dimension $s \times r$ with the (i, j)-th entry given by $\dfrac{\partial f_j}{\partial u_i}$.

Lemma: The derivative matrix g' ($(2p + m) \times p$) is given by,

$$g'\left(\hat{\beta}_{ML}, \hat{\theta}_{ML}, \hat{\beta}^0_{ML}\right) = [\{I - \hat{V}_{\hat{\beta}_{ML}}(\hat{V}_{\hat{\beta}_{ML}} + \hat{\Delta}^\top\hat{\theta}_{ML}\hat{\theta}^\top_{ML}\hat{\Delta})^{-1}\}, E^\top, \{\hat{V}_{\hat{\beta}_{ML}}(\hat{V}_{\hat{\beta}_{ML}} + \hat{\Delta}^\top\hat{\theta}_{ML}\hat{\theta}^\top_{ML}\hat{\Delta})^{-1}\}]^\top.$$

where, E is a $m \times p$ matrix,

$$\begin{aligned}E &= \frac{\hat{\theta}^\top_{ML}\hat{\Delta}(\hat{V}_{\hat{\beta}_{ML}})^{-1}(\hat{\beta}_{ML} - \hat{\beta}^0_{ML})\hat{\Delta} + \hat{\Delta}(\hat{V}_{\hat{\beta}_{ML}})^{-1}(\hat{\beta}^0_{ML} - \hat{\beta}_{ML})\hat{\theta}^\top_{ML}\hat{\Delta}}{\{1 + \hat{\theta}^\top_{ML}\hat{\Delta}(\hat{V}_{\hat{\beta}_{ML}})^{-1}\hat{\Delta}^\top\hat{\theta}_{ML}\}} \\ &- \frac{2\hat{\theta}^\top_{ML}\hat{\Delta}(\hat{V}_{\hat{\beta}_{ML}})^{-1}(\hat{\beta}_{ML} - \hat{\beta}^0_{ML})\hat{\Delta}(\hat{V}_{\hat{\beta}_{ML}})^{-1}\hat{\Delta}^\top\hat{\theta}_{ML}\hat{\theta}^\top_{ML}\hat{\Delta}}{\{1 + \hat{\theta}^\top_{ML}\Delta(\hat{V}_{\hat{\beta}_{ML}})^{-1}\hat{\Delta}^\top\hat{\theta}_{ML}\}^2}\end{aligned}$$

.

Proof: The above expression follows by first noticing that the matrix $(\hat{V}_{\hat{\beta}_{ML}} + \Delta^\top\hat{\theta}_{ML}\hat{\theta}^\top_{ML}\Delta)^{-1}$ is of the form $(V + uu^\top)^{-1}$ which could be expanded as in (i) below. Further simplification is carried out by noticing that, for any matrix A and column vectors u and c, of conformable multiplication

orders as needed, the derivative of the vector $Auu'c$ with respect to u can be obtained as in (ii).

$$\begin{aligned} \text{(i). } (V + uu^\top)^{-1} &= V^{-1} - \frac{(V^{-1}u)(u^\top V^{-1})}{1 + u^\top V^{-1}u} \\ \text{and, (ii). } \frac{\partial}{\partial u} Auu^\top c &= u^\top cA + cu^\top A. \end{aligned}$$

The final expression in the lemma is obtained by first using (i), on the expression for $\hat{\beta}_{EB}$ as in (12), followed by using the quotient rule of differentiation. The result in (ii) is used to differentiate terms of the form $A\Delta^\top \hat{\theta}_{ML}\hat{\theta}_{ML}^\top \Delta c$ with respect to $\hat{\theta}_{ML}$ which appear in the numerator of the quotient. Some algebraic manipulation leads to the expression for E.

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

Table 2: Simulation results showing mean squared error 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 $\beta = \log(\psi) = \log(2) = 0.6931$.

		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)
	two-stage	0.26 (0.00)	0.13(0.01)	0.05 (0.00)
θ_{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)
	two-stage	0.31 (0.16)	0.19 (0.15)	0.10 (0.13)
$\theta_{GE} = \log(1.5)$	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)
	two-stage	0.44 (0.27)	0.28 (0.22)	0.15 (0.13)
$\theta_{GE} = \log(2)$	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)	0.26(0.20)	0.11 (0.12)
	two-stage	0.67 (0.34)	0.38 (0.19)	0.11 (0.03)

Table 3: Odds ratios of association between NAT2 acetylation activity (rapid vs slow) and smoking (ever vs. never) estimated using controls from 11 different case-control studies (Marcus et al., 2000)

Study site	OR_{GE}	$\theta_{GE} = \log(OR_{GE})$	P-value for testing $H_0 : \theta_{GE} = 0$
Brockmoller	1.16	0.15	0.55
Taylor	1.44	0.37	0.21
Risch	1.85	0.62	0.39
Mommsen	0.88	-0.13	0.81
Roots	0.56	-0.57	0.29
Kaisary	0.76	-0.28	0.48
Romkes	0.95	-0.06	0.92
Dewan	1.04	0.03	0.94
Ishizu	1.23	0.21	0.73
Karakaya	1.39	0.33	0.42

Table 4: Simulation results showing mean-squared-error and bias (in parentheses) in estimation of the interaction parameter $\beta = \log(\psi)$ when the control odds ratio between G and E, namely, θ_{GE} is generated from a N(0.07,0.35) distribution in each run, reflecting the distribution elicited from Marcus et al., 2000 meta analysis. The prevalences of G and E are fixed at $P_G = P_E = 0.3$. The parameters in the disease risk model are fixed at $OR_{10} = OR_{01} = 1$ and $\beta = \log(\psi) = \log(2) = 0.6931$.

	Sample Size		
	$n_0 = n_1 = 100$	$n_0 = n_1 = 200$	$n_0 = n_1 = 500$
case-control	0.45 (0.03)	0.22 (0.01)	0.08 (0.00)
case-only	0.33 (0.08)	0.23 (0.07)	0.17 (0.06)
EB	0.33 (0.05)	0.18 (0.03)	0.08 (0.02)
EB-TRUE ($\tau = 0.35$)[1]	0.30 (0.06)	0.17 (0.04)	0.08 (0.03)

[1]: This is the Bayes estimate obtained by using the true simulation value of the prior parameter τ instead of estimating τ from the marginal likelihood.

Table 5: Simulation results showing mean-squared error 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: $P(E_1 = 1) = 0.3, P(E_2 = 1) = 0.3, OR_{E_1\ E_2} = 2.0$, $P(G = 1) = 0.3, OR_{GE_1} = 1, OR_{GE_2} = 1.5$. The parameters for the disease risk model were set at $\beta_G = \beta_{E_1} = \beta_{E_2} = 0$, and, $\beta_{G*E_1} = \beta_{G*E_2} = \log(2)$.

		$MSE1$ $(G * E_1)$	$MSE2$ $(G * E_2)$	$MSE1$ $+MSE2$
$n_0 = n_1 = 100$	Dependence	0.46 (0.04)	0.48 (0.10)	0.94
	Independence	0.20 (0.04)	0.39 (0.44)	0.59
	Empirical Bayes	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
	Empirical Bayes	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
	Empirical Bayes	0.06 (0.00)	0.09 (0.12)	0.15

Table 6: Analysis of Israeli ovarian cancer data: Estimates of the log odds-ratio parameters corresponding to each effect is provided, accompanied with 95% confidence intervals.[1]

	$\hat{\beta}_{BRCA1/2}$	$\hat{\beta}_{OC}$	$\hat{\beta}_{Parity}$	$\hat{\beta}_{G*OC}$	$\hat{\beta}_{G*Parity}$
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)
Empirical Bayes	3.270	-0.051	-0.062	0.070	-0.077
CI	(2.418, 4.121)	(-0.108,0.006)	(-0.127,0.005)	(0.002,0.139)	(-0.185,0.032)

[1] The analysis is adjusted for effects of age, ethnicity, personal history of breast cancer, family history of breast or ovarian cancer and history of gynaecological surgery.

Table 7: Analysis of colorectal adenoma data: Estimates of the log odds-ratio parameters corresponding to each effect is provided, accompanied with 95% confidence intervals.[2]

	$\hat{\beta}_{NAT2=2}$	$\hat{\beta}_{SMK=1}$	$\hat{\beta}_{SMK=2}$	$\hat{\beta}_{NAT2=2*SMK=1}$	$\hat{\beta}_{NAT2=2*SMK=2}$
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)
Empirical Bayes	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)

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

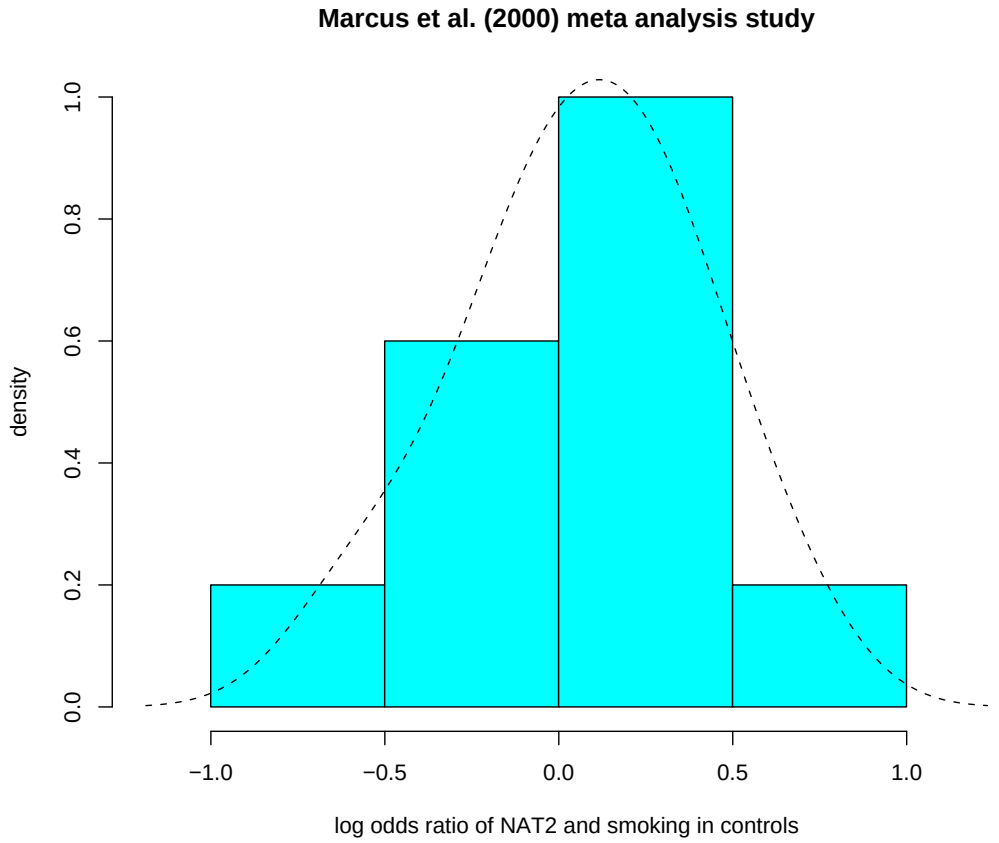

Figure 1: Histogram of log odds-ratios between NAT2 and smoking in control subjects from 11 case-control studies conducted in several study sites across US, Europe, Japan and India (Marcus et al., 2000).