Biometrics 66, 934–948
September 2010

DOI: 10.1111/j.1541-0420.2009.01357.x

Case–Control Studies of Gene–Environment Interaction: Bayesian Design and Analysis

Bhramar Mukherjee,[1,*] Jaeil Ahn,[1] Stephen B. Gruber,[2] Malay Ghosh,[3] and Nilanjan Chatterjee[4]

[1]Department of Biostatistics, University of Michigan
[2]Department of Epidemiology, Human Genetics and Internal Medicine, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[3]Department of Statistics, University of Florida, Gainesville, Florida 32611, U.S.A.
[4]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland 20852, U.S.A.
[*]*email:* bhramar@umich.edu

Summary. With increasing frequency, epidemiologic studies are addressing hypotheses regarding gene-environment interaction. In many well-studied candidate genes and for standard dietary and behavioral epidemiologic exposures, there is often substantial *prior* information available that may be used to analyze current data as well as for designing a new study. In this article, first, we propose a proper full Bayesian approach for analyzing studies of gene–environment interaction. The Bayesian approach provides a natural way to incorporate uncertainties around the assumption of gene–environment independence, often used in such an analysis. We then consider Bayesian sample size determination criteria for both estimation and hypothesis testing regarding the multiplicative gene–environment interaction parameter. We illustrate our proposed methods using data from a large ongoing case–control study of colorectal cancer investigating the interaction of N-acetyl transferase type 2 (NAT2) with smoking and red meat consumption. We use the existing data to elicit a design prior and show how to use this information in allocating cases and controls in planning a future study that investigates the same interaction parameters. The Bayesian design and analysis strategies are compared with their corresponding frequentist counterparts.

Key words: Case-only design; Gene–environment independence; Highest posterior density interval; Molecular epidemiology of colorectal cancer; Multinomial-Dirichlet; Posterior odds.

1. Introduction

Case–control (CC) studies are popular epidemiological tools for assessing association between a disease and a candidate gene. Most human diseases have a multifactorial etiology, involving complex interplay of multiple genetic and environmental factors. Thus, while searching for the disease-causing variant(s), one cannot ignore the environmental risk factors modifying the disease risk. Similarly, one has to understand the genetic architecture of a disease while evaluating risk due to standard environmental exposures such as environmental toxins, dietary exposures, and physical activity levels. The National Institutes of Health has recently undertaken the Genes, Environment and Health Initiative (GEI: `http://www.gei.nih.gov`) to integrate genomic sciences with the world of assessing environmental exposures in researching disease etiology. With these new initiatives, the emphasis on searching for gene–environment interaction (G × E) effects is becoming more common and fundamental in determining the genetic and environmental roots of complex diseases.

An important issue in estimating G × E interaction is the use of an assumption that there is no association between the genetic factor and the environmental factor (gene–environment independence assumption; Piegorsch, Weinberg, and Taylor, 1994). Under this assumption, the multiplicative interaction odds-ratio (OR) parameter can be estimated by data on cases only. Exploiting this assumption by means of retrospective likelihoods leads to an enormous gain in efficiency for estimating the G × E interaction parameter in a general regression model (Chatterjee and Carroll, 2005). However, the methods using gene–environment independence assumption incur bias and result in inflated Type-I error rates under departures from this assumption (Albert et al., 2001; Mukherjee et al., 2008). There have been recent proposals on relaxing the gene–environment independence assumption in an empirical Bayes (EB) fashion (Mukherjee and Chatterjee, 2008) or through Bayesian model averaging (BMA; Li and Conti, 2009). Several simulation studies in the above papers illustrate that under violation of the independence assumption, one can trade off between bias and efficiency in a data-adaptive way by using such composite estimators. In the current article, we propose a full Bayes (FB) analysis and design strategy to incorporate prior belief on the assumption of gene–environment independence. Unlike the EB or BMA approach, the FB approach retains the advantage of proper Bayesian inference based on the exact posterior, without relying on the validity of the large sample Wald tests based on the asymptotic distribution of the EB or the BMA estimator. We derive a closed-form expression for the posterior distribution of the interaction OR in terms of standard Beta random variables and thus there is no additional computational burden associated with the FB approach. Based on the new FB analysis strategy devised in the first half of the article, we

proceed to evaluate sample size considerations from a proper Bayesian perspective.

There has been considerable amount of work on the sample size determination (SSD) problem for unmatched CC studies of G × E interaction (Hwang et al., 1994; Foppa and Spiegelman, 1997; Goldstein et al., 1997; Garcia-Closas and Lubin, 1999; Luan, Wong, and Day, 2001). Yang, Khoury, and Flanders (1997) discuss the SSD problem for case-only (CO) designs. Gauderman (2002) considered the SSD problem under several study designs and offered a popular software, called `QUANTO`, widely used by many investigators (`http://hydra.usc.edu/gxe`). Mukherjee et al. (2008) evaluate the newly proposed shrinkage estimators of G × E interaction in terms of power and Type-I error in unmatched CC studies. Another genre of closely related literature where sample sizes are calculated for indirect use of G × E interaction in association studies for detecting genetic main effects (Kraft et al., 2007; Hein, Beckmann, and Chang-Claude, 2008) has emerged, and the proposed Bayesian SSD methods could easily be adapted to the latter context.

In all of the above papers, the design approach followed is purely frequentist, where, in the planning stages of a study that aims at testing significance of the interaction OR ψ (say), certain characteristics of the *sampling distribution* of the test statistic (type-I error, power, length of confidence interval) are controlled. Since the criteria typically involve unknown parameter values, for example, prevalence of G and E, OR corresponding to main effects of G and E, OR between G and E in the control population, and a plausible value for the interaction parameter ψ, initial guesses for the true parameters are needed for evaluation and implementation of these procedures. In this sense, the SSD problem has a natural Bayesian flavor even in a frequentist setting as it requires some form of prior information. The resulting sample sizes are only *locally optimal* choices and depend on these initial guesses that are often overly optimistic or potentially incorrect. Bayesian methods do not suffer from this limitation and one can allow different levels of uncertainty/information on the design parameters via prior distributions.

Bayesian SSD methods control certain aspects of the *posterior distribution*, such as precision of posterior estimates, length of Bayesian credible intervals, and alike. These "characteristics" of the posterior distributions are functionals of data, and thus are random quantities under a pre-experimental setting. The probability distribution governing these random functionals is the prior predictive distribution of data. The prior predictive distribution is the marginal distribution of data integrated with respect to the prior distribution. Thus, the resultant sample sizes that control for these criteria depend on the *entire* prior distribution, not just fixed point guesses for the parameters. Another less-critical limitation of many of the frequentist sample size methods is their dependence on asymptotic distribution of the maximum likelihood estimate (MLE), whereas Bayesian methods rely on exact pre-posterior computations.

The literature on Bayesian SSD has been flourishing lately. An ensemble of papers on Bayesian SSD had appeared in *The Statistician* from 1995 to 1997 (Adcock, 1997; Bernardo, 1997; Joseph, du Berger, and Bélisle, 1997; Lindley, 1997; Pham-Gia, 1997; Joseph and Wolfson, 1997, to name a few). Wang and Gelfand (2002) consider Monte Carlo methods for the Bayesian SSD problem whereas Walker (2003) presents a Bayesian nonparametric alternative to the SSD problem. More recently, De Santis and his colleagues have made a series of important new contributions to the field (De Santis, 2007, 2006), with the first Bayesian SSD paper on CC studies by De Santis, Pacifico, and Sambucini in 2004. M'Lan, Lawrence, and Wolfson (2006) propose Monte Carlo methods for Bayesian SSD and propose several criteria related to interval estimation of the OR parameter in a CC study. In a section of their paper, M'Lan et al. (2006) consider the Bayesian SSD problem for G × E studies, which, to the best of our knowledge, is so far the only published work on Bayesian SSD for G × E problems. M'Lan et al. (2006) consider an array of Bayesian SSD criteria, study them analytically, and propose elegant computational tools to identify the optimal SSD. Though M'Lan et al. (2006) do consider Bayesian SSD for CO and CC analysis of G × E studies, the paper does not address the particular issue of incorporating prior belief on gene–environment independence. M'Lan et al. (2006) consider Bayesian SSD criteria related to interval estimation whereas we consider criteria related to testing as well. The current article attempts to present a more complete and comprehensive account of Bayesian analysis and SSD for G × E problems, elaborates on a real application, thus adding to the timely first presentation in M'Lan et al. (2006).

The rest of the paper is organized as follows. Section 2 presents a description of the FB approach for G × E studies. We use data on NAT2, smoking, and red meat consumption from the molecular epidemiology of colorectal cancer (MECC) study to illustrate the analytical methods. We address the problem of relaxing gene–environment independence in Section 2.1 and compare the CC, CO, EB, and the newly proposed FB estimator in terms of bias and mean squared error (MSE). Section 3 describes the different Bayesian SSD criteria we consider. Section 4 contains a description of the general method to determine optimal allocation of cases and controls and the computational algorithm. Section 5 uses the analysis results obtained in Section 2 to plan a new study and elicit design priors based on the current results. This section then provides sample size choices under different criteria that we consider. Sensitivity of the allocation rule with respect to prior strength is assessed. We evaluate the Bayesian SSD criteria under corresponding frequentist case–control designs that are routinely used. Section 6 presents concluding discussion while some algebraic and numerical details are relegated to the Web Appendix.

2. Bayesian Analysis of Gene–Environment Interaction

Bayesian analysis of CC data has recently received much attention (Müller and Roeder, 1997; Gustafson, Le, and Vallee, 2002; Sinha, Mukherjee, and Ghosh, 2004; Sinha et al., 2005; Mukherjee et al., 2007). Lindley (1964) remains a classic reference on Bayesian analysis of contingency tables. Latorre (1982, 1984) consider deriving analytical expression for posterior inference regarding cross-ratio estimation in 2×2 and 2×4 tables. We propose the following simple Bayesian

Table 1
Data structure for a unmatched CC 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

analysis of a 2×4 table for the situation with a binary G and a binary E and the representation of the posterior distribution of the interaction OR parameter. The results can be directly extended to categorical G and E.

We consider the set-up of an unmatched CC 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 $\boldsymbol{r}_0 = (r_{000}, r_{001}, r_{010}, r_{011})$ and $\boldsymbol{r}_1 = (r_{100}, r_{101}, r_{110}, r_{111})$ denote the vector of observed cell frequencies in the controls and cases, respectively. The population parameters, namely, the cell probabilities corresponding to a particular G–E configuration in the underlying control and case populations are denoted as $\boldsymbol{p}_0 = (p_{000}, p_{001}, p_{010}, p_{011} = 1 - p_{000} - p_{001} - p_{010})$ and $\boldsymbol{p}_1 = (p_{100}, p_{101}, p_{110}, p_{111} = 1 - p_{100} - p_{101} - p_{110})$, respectively. The observed vectors of cell counts can be viewed as realizations from two independent multinomial distributions, namely, $\boldsymbol{r}_0 \,|\, \boldsymbol{p}_0 \sim \text{Multinomial}(n_0, \boldsymbol{p}_0)$ and $\boldsymbol{r}_1 \,|\, \boldsymbol{p}_1 \sim \text{Multinomial}(n_1, \boldsymbol{p}_1)$. Let $OR_{10} = p_{000}p_{101}/p_{001}p_{100}$ denote the OR associated with E for nonsusceptible subjects $(G = 0)$, $OR_{01} = p_{000}p_{110}/p_{010}p_{100}$ denote the OR associated with G for unexposed subjects $(E = 0)$ and $OR_{11} = p_{000}p_{111}/p_{011}p_{100}$ denote the OR 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. Also note that

$$\psi = \frac{p_{100}p_{111}}{p_{101}p_{110}} \Big/ \frac{p_{000}p_{011}}{p_{001}p_{010}} = \psi_1/\psi_0 = \frac{\text{GE OR in cases}}{\text{GE OR in controls}}. \quad (1)$$

In (1), ψ_0 is the gene–environment association parameter that is assumed to be unity under G–E independence assumption. The likelihood of the parameters given the data is then a product of two independent multinomial likelihoods with probability vectors $\boldsymbol{p}_d$, $d = 0, 1$. We assume independent Dirichlet prior distribution on $\boldsymbol{p}_d$, namely, $\pi(\boldsymbol{p}_d) \sim \text{Dirichlet}(\boldsymbol{\alpha}_d = (\alpha_{d00}, \alpha_{d01}, \alpha_{d10}, \alpha_{d11}))$, $d = 0, 1$. One can derive the posterior distribution of $\boldsymbol{p}_d$ again as a Dirichlet distribution via this multinomial-Dirichlet conjugate analysis.

$$p(\boldsymbol{r}_d \,|\, \boldsymbol{p}_d) \sim \text{Multinomial}(n_d, \boldsymbol{p}_d);$$
$$\pi(\boldsymbol{p}_d) \sim \text{Dirichlet}(\boldsymbol{\alpha}_d);$$
$$\Rightarrow \pi(\boldsymbol{p}_d \,|\, \boldsymbol{r}_d) \sim \text{Dirichlet}(r_{d00} + \alpha_{d00}, r_{d01} + \alpha_{d01}, r_{d10} + \alpha_{d10}, r_{d11} + \alpha_{d11}). \quad (2)$$

Result 1: Let us consider the following transformation from $\boldsymbol{p}_d \to (X_d, Y_d, Z_d)$, $d = 0, 1$.

$$Z_d = p_{d00}, \quad Y_d = \frac{p_{d01}}{1 - p_{d00}}, \quad X_d = \frac{p_{d10}}{1 - p_{d00} - p_{d01}}.$$

Then, one can show that the posterior distributions

$$Z_d \,|\, \boldsymbol{r}_d \sim \text{Beta}(r_{d00} + \alpha_{d00}, r_{d01} + \alpha_{d01} + r_{d10} + \alpha_{d10} + r_{d11} + \alpha_{d11}),$$
$$Y_d \,|\, \boldsymbol{r}_d \sim \text{Beta}(r_{d01} + \alpha_{d01}, r_{d10} + \alpha_{d10} + r_{d11} + \alpha_{d11}),$$
$$X_d \,|\, \boldsymbol{r}_d \sim \text{Beta}(r_{d10} + \alpha_{d10}, r_{d11} + \alpha_{d11}).$$

And $(X_d, Y_d, Z_d \,|\, \boldsymbol{r}_d)$ are mutually independent random variables.

Proof. Follows by elementary Jacobian calculation corresponding to the transformation. For a similar argument for cross-ratio estimation in a 2×2 table, see Lecoutre and Charron (2000). Note that we can express the GE OR parameter in cases and controls as, (for $d = 0, 1,$)

$$\psi_d = \frac{Z_d(1 - X_d)}{X_d Y_d(1 - Z_d)}. \quad (3)$$

Finally, the interaction log OR can be expressed as $\beta = \log(\psi) = \log(\psi_1) - \log(\psi_0)$.

We assume that the stochastic model parameters in cases and controls, namely, $\boldsymbol{p}_1$ and $\boldsymbol{p}_0$ are independent, so $\pi(\boldsymbol{p}_0, \boldsymbol{p}_1) = \pi(\boldsymbol{p}_0)\pi(\boldsymbol{p}_1)$. Thus, in the proposed FB analysis with the above prior-likelihood-posterior structure, it is extremely easy to generate the posterior distribution of β, namely $\pi(\beta \,|\, \boldsymbol{r}_0, \boldsymbol{r}_1)$ in terms of $(X_d, Y_d, Z_d \,|\, \boldsymbol{r}_d)$ which are independent Beta random variables and one does not have to resort to any numerical integration techniques. This, we believe, is a very attractive feature of the FB approach as it provides the flexibility of obtaining the full posterior distribution of the interaction parameter using extremely inexpensive computation. Latorre (1984) derived expression for the posterior distribution of ψ using sum of four infinite series, but we work with exact simulation from the posterior distribution of ψ instead of using these expressions.

Remark 1*: Asymptotic normality of the posterior.* The asymptotic normality of the posterior distribution follows by approximating the joint posterior distribution of $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ by Laplace approximation around the posterior mode (or the MLE; Ghosh, Delampady, and Samanta, 2006, p. 62). Multivariate delta theorem can then be used to derive the asymptotic approximation for the distribution of $\log(\psi_0)$ and $\log(\psi_1)$. By independence of the data and prior parameters in cases and controls, the asymptotic posterior distribution for $\log(\beta)$ is derived as,

$$\log(\beta) \,|\, \boldsymbol{r}_0, \boldsymbol{r}_1 \xrightarrow{d} N\big(\mu_1 - \mu_0, \sigma_0^2 + \sigma_1^2\big),$$

where, for $d, g, e = 0, 1$,

$$\mu_d = \log\left(\frac{\tilde{p}_{d00}\tilde{p}_{d11}}{\tilde{p}_{d01}\tilde{p}_{d10}}\right), \qquad \sigma_d^2 = \sum_{g=0}^{1}\sum_{e=0}^{1}(1/\tilde{p}_{dge}),$$

$$\tilde{p}_{dge} = \frac{r_{dge} + \alpha_{dge}}{n_d + \sum_{g=0}^{1}\sum_{e=0}^{1}\alpha_{dge}}.$$

For approximation around the MLE one can simply replace $\tilde{p}_{dge}$ by $\hat{p}_{dge} = r_{dge}/n_d$. Again, though sample size calculations in terms of a normal variate may be convenient, we do not use this approximation in our subsequent calculations as generation from the exact posterior distribution itself is straightforward.

Remark 2: Prior structure. Instead of the cell probability vectors $(\boldsymbol{p}_0, \boldsymbol{p}_1)$, from an epidemiologic perspective, it may be more natural to have prior guesses on an alternative parametrization, say for example in terms of $(P_G, P_E, OR_{01}, OR_{10}, \psi_0, \psi)$. We have formulated the problem in terms of the multinomial-Dirichlet structure for computational convenience. However, one can use the following relationships to translate the prior guesses back and forth between the two parameterizations. Directly imposing normal priors on the log-OR scale will distort the conjugacy of the analysis that we would like to retain.

The control probability vector $\boldsymbol{p}_0$ is related to (P_G, P_E, ψ_0), by the following equation:

$$\psi_0 = \frac{p_{000}(p_{000} - (1 - P_G - P_E))}{(1 - P_G - p_{000})(1 - P_E - p_{000})},$$
$$p_{001} = 1 - P_G - p_{000}, \qquad p_{010} = 1 - P_E - p_{000}.$$

The case-probability vector $\boldsymbol{p}_1$ can be obtained from the disease risk parameters and $\boldsymbol{p}_0$ as

$$p_{100} = \frac{p_{000}}{p}, \qquad p_{101} = \frac{p_{001}}{p}\cdot OR_{10},$$
$$p_{110} = \frac{p_{010}}{p}\cdot OR_{01}, \qquad p_{111} = \frac{p_{011}}{p}\cdot OR_{10}\cdot OR_{01}\cdot\psi,$$

where $p = p_{000} + p_{001}\cdot OR_{10} + p_{010}\cdot OR_{01} + p_{011}\cdot OR_{10}\cdot OR_{01}\cdot\psi$.

In the next section, we present FB inference regarding the interaction log OR parameter β, through our analysis of the MECC data. We specially focus on the issue of incorporating uncertainty around the gene–environment (G–E) independence assumption through our prior specification on the *G–E* association parameter $\log(\psi_0)$.

2.1 *Incorporating Uncertainties Around G–E Independence*

In the 2×4 Table 1, the unconstrained MLE of the interaction parameter β based on CC data is given by

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

The *G–E* independence assumption is realistic for external exposures such as environmental pollutants, toxins, or radioactive substances, but is at best dubious for genes and behavioral exposures that may share a common metabolic pathway. Note that under the assumption of *G–E* independence, the denominator in (1), namely $\psi_0 \equiv 1$, and one can unbiasedly estimate $\beta = \log(\psi)$ by data on cases only, namely, by the CO estimator of β,

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

This CO estimator is much more efficient than $\hat{\beta}_{CC}$, but incurs bias when $\psi_0 \neq 1$.

Mukherjee et al. (2008) present a comprehensive discussion of different estimators of G × E interaction parameter including an EB estimator that combines $\hat{\beta}_{CC}$ and $\hat{\beta}_{CO}$ in a data-adaptive way, depending on the strength of *G–E* association. The FB analysis proposed in this article provides another option to incorporate uncertainty around *G–E* independence and still provides shrinkage when the assumption is tenable in the light of the data. We work with a completely general CC likelihood that does not impose any assumption on *G–E* independence. The independence assumption $(\log(\psi_0) = 0)$ is reflected only through the prior structure. We assume a prior on the *G–E* association parameter $\log(\psi_0)$ that is centered around zero to reflect *G–E* independence but vary the strength/variance of the prior to allow uncertainty around the assumption of independence. The advantage of this approach over the EB approach is to retain the flexibility of a full Bayesian inference and obtain the entire posterior of data. Moreover, the FB approach is based on exact computations, whereas the variance expression and derived Wald tests for the EB approach depend on the accuracy of certain large sample approximations and asymptotic normality of the sampling distribution of the EB estimator. The limitation of the FB approach is sensitivity of the results to prior choices and how to select the prior strength/uncertainty.

We first define the strength of the Dirichlet prior on the control and case probability vector as $s_d = \sum_{g=0}^{1}\sum_{e=0}^{1}\alpha_{dge}$ for $d = 0, 1$. The total prior strength is given by $s = s_0 + s_1$. One can choose s_0 and s_1 based on the level of uncertainty of the data coming from the control and case population respectively. For example, under the same level of uncertainty in both the case and control parameters, and a fairly noninformative prior choice that centers the prior on β roughly around zero, one may select $\boldsymbol{\alpha}_0 = \boldsymbol{\alpha}_1 = (5, 5, 5, 5)$ implying a total prior strength of $s = s_0 + s_1 = 20 + 20 = 40$. Different choices of s_0 and s_1 induce different variances on the parameters $\log(\psi_0)$ and $\log(\psi_1)$, an expression is provided in the Web Appendix. For reflecting different degrees of belief on *G–E* independence assumption, we treat $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ asymmetrically in our prior specification. We choose a fairly noninformative prior on $\boldsymbol{p}_1$, and we choose a prior on $\boldsymbol{p}_0$ that leads to a prior on $\log(\psi_0)$ roughly centered at zero (or *G–E* independence). We then vary s_0, to induce different prior variances around the independence assumption.

We now describe FB analysis of the MECC study data with different choices of s_0. The MECC study is a population-based CC study of patients who received a diagnosis of invasive colorectal cancer (CRC) in northern Israel between March 31, 1998, and March 31, 2004. Participants were interviewed to obtain demographic information, personal and family history of cancer, medical history, medication use, and health habits. They also completed a dietary questionnaire and a blood sample was collected. NAT2 genotyping was used to classify individuals as fast or slow acetylators for consistency with

previously published literature (Roberts-Thomson et al., 1996). The acetylator phenotype has been hypothesized to modulate the relationship between red meat and risk of CRC, with increasing red meat consumption associated with increased risk of CRC among fast, but not slow acetylators in some (Roberts-Thomson et al., 1996), but not all studies (Barrett et al., 2003). NAT2 has also been suggested as a potential modifier of the relationship between cigarette smoking and CRC as well as colorectal adenomas, although studies are not consistent (Barrett et al., 2003). We analyze the interaction between NAT2 and frequency of grilled red meat consumption (RMGF) as well as NAT2 and smoking status (SMOKE) with the existing study base of 1785 cases and 1971 controls. The observed cell counts are presented in Web Table 1 for NAT2, SMOKE, and RMGF, stratified by CC status.

Table 2 presents the dataset from the MECC study analyzed with the three different methods (CC, CO, and EB), and the proposed FB approach with varying levels of prior variance on $\log(\psi_0)$, whereas the prior parameter on $\boldsymbol{p}_1$, namely, $\boldsymbol{\alpha}_1$ is fixed at a fairly noninformative choice (5, 5, 5, 5). One can notice that for NAT2 $*$ RMGF, the CC estimator of β is 0.39, whereas the CO estimator is 0.11. The MLE of $\log(\psi_0)$ is -0.27 with SE=0.14. By varying the prior strength on $\boldsymbol{p}_0$ from 20 to 320 (inducing a change in prior variance on $\log(\psi_0)$ from 0.9 to 0.05), one can trade off between the CC and CO analysis with FB estimates varying from 0.38 to 0.13 respectively. For NAT2 $\times$ SMOKE, the CC estimator is 0.18 whereas the CO estimator is -0.12. In contrast to the situation with RMGF, for SMOKE, even with a very sharp prior around zero with strength 320, reflecting substantial prior belief in *G–E* independence, the FB inference tracks the CC estimator (FB estimate with prior strength 320 is 0.11). This is because the data contain quite strong evidence that $\log(\psi_0)$ departs from 0, that NAT2 and SMOKE exhibit departures from independence (MLE of $\log(\psi_0)$ is -0.30, SE = 0.09). Thus unlike the NAT2 $\times$ RMGF data, the NAT2 $\times$ SMOKE data resist the prior assumption of $G-E$ independence and cannot be shrunk toward a CO analysis even with a quite strong prior around this assumption. The EB estimator (Mukherjee and Chatterjee, 2008) in Table 2 also trades off between the CC and CO estimators depending on the strength of *G–E* independence reflected in the data. In Web Table 2, we present posterior estimates of $\log(\psi_1)$ and $\log(\psi_0)$ separately, to assess effects of the prior. Figure 1 illustrates the posterior of β under three different prior uncertainties on *G–E* independence with the CC, CO, and EB estimates marked by arrows on the horizontal axis. One can notice the changes in posterior behavior corresponding to these two different exposures reflecting different levels of evidence in support of the independence assumption in the two datasets. The differences in the CC and CO estimates in both datasets reflect how drastically the assumption of *G–E* independence can affect inference regarding G $\times$ E interaction. A supplementary analysis that treats $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ symmetrically is presented in Web Table 3 and Web Figure 1.

Table 2

The different estimates of log-interaction parameter β in the MECC data corresponding to NAT2 $\times$ RMGF and NAT2 $\times$ SMOKE interaction. CC, CO, and EB refer to the case–control, case-only, and empirical Bayes estimator of β while under the FB analysis, we use a fairly noninformative prior on the G–E log OR in cases, namely, $\log(\psi_1)$, and vary the strength of the prior s_0 on $G-E$ log OR among controls, i.e., $\log(\psi_0)$, to reflect different levels of uncertainty on gene–environment independence. For FB analysis "Est" refers to the posterior mean whereas "PSD" is the posterior standard deviation.

		NAT2 $\times$ RMGF		NAT2 $\times$ SMOKE	
Method		Est.	SE	Est.	SE
CC		0.394	0.194	0.177	0.137
CO		0.114	0.132	−0.121	0.099
EB		0.303	0.206	0.125	0.145
	s_0	Est	PSD	Est	PSD
FB†	20	0.377	0.189	0.175	0.137
	80	0.304	0.178	0.165	0.135
	320	0.130	0.174	0.112	0.132

†For FB, the case Dirichlet parameter α_1 is fixed as (5, 5, 5, 5) with $s_1 = 20$ whereas the prior strength on controls namely s_0 of 20, 80, and 320 correspond to the control Dirichlet parameter α_0 being set at (5, 5, 5, 5), (20, 20, 20, 20), and (80, 80, 80, 80) respectively. Prior variance corresponding to $\log(\psi_1)$ is 0.90 whereas prior variance on $\log(\psi_0)$ under $s_0 = 20$, 80, and 320 are 0.90, 0.20, and 0.05 respectively.

Table 3 presents results of a small-scale simulation study comparing the CC, CO, EB, and the FB estimate with varying prior strength on $\log(\psi_0)$. We consider four simulation settings, (1) when independence assumption holds with $\psi_0 = 1$; (2) under modest departure from the independence assumption, with positive *G–E* association, $\psi_0 = 1.25$. Settings (3) and (4) are based on the observed NAT2 $\times$ RMGF and NAT2 $\times$ SMOKE configurations that exhibit negative association between *G* and *E*. Under the independence assumption in (1) FB 80 and FB 320 have significant MSE advantage over the classical CC estimator for modest sample sizes, the gain is comparable to the CO estimator. Under violation of the independence assumption, FB 20 and FB 80 provide a better control on the bias when compared to the CO estimator. FB 80 often exhibits reduced bias compared to the EB estimator with similar or less MSE values. The trade-off between bias and MSE of the FB approach depends on the strength of the prior. Though the FB prior strength regarding *G–E* association can be chosen in a data adaptive way, resembling the EB approach, or be elicited from historical data, using a generic prior of moderate strength like FB 80 appears to be a fairly robust choice across all sample sizes and *G–E* association scenarios for this example.

3. SSD Criteria

The criteria for SSD in the Bayesian domain stem mainly from two different perspectives. One approach as introduced by Raiffa and Schlaifer (1961) and more recently by Bernardo (1997) and Lindley (1997) is to pose the problem in a pure utility based decision theoretic framework. The other approach (which may not always reconcile with a maximizing expected utility principle) followed by Joseph, Wolfson, and du Berger (1995); Pham-Gia (1997); and De Santis (2007), is to consider a functional of the posterior

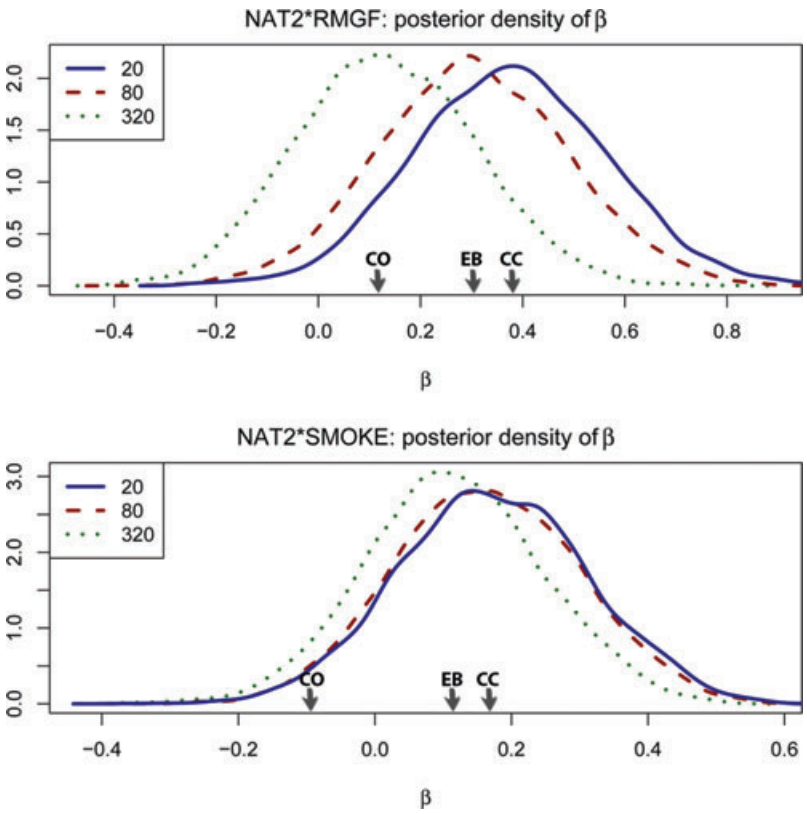

Figure 1. Posterior density corresponding to NAT2 × RMGF (top) and NAT2 × SMOKE (bottom) interaction log OR under three different levels of uncertainty around gene–environment independence. In each case, the Dirichlet prior parameter on $\boldsymbol{p}_1$ is kept fixed at $\boldsymbol{\alpha}_1 = (5, 5, 5, 5)$ and the prior parameter on $\boldsymbol{p}_0$, namely, $\boldsymbol{\alpha}_0$, is varied at (5, 5, 5, 5), (20, 20, 20, 20), and (80, 80, 80, 80) corresponding to the three posteriors denoted by s_0 set at 20, 80, and 320 in the above figure. The arrows on the horizontal axis mark the values of the case–control (CC), the case-only (CO), and the empirical Bayes (EB) estimate. The corresponding numerical results are collected in Table 2. This figure appears in color in the electronic version of this article.

distribution, routinely used in Bayesian inference. Recall that at the design stage, before one observes data, functionals of the posterior distribution are random variables that are governed by the prior predictive density. For the 2×4 table described in Section 2, the prior predictive distribution of the random cell counts $\boldsymbol{R}_d$ under the data and prior structure of (2) is given by the following multinomial-Dirichlet distribution, for $d = 0, 1$,

$$\begin{aligned}
\mathbb{P}(\boldsymbol{R}_d = \boldsymbol{r}_d) &= \int p(\boldsymbol{R}_d = \boldsymbol{r}_d \mid \boldsymbol{p}_d)\pi(\boldsymbol{p}_d)d\boldsymbol{p}_d \\
&= \binom{n_d}{r_{d00}, r_{d01}, r_{d10}, r_{d11}} \frac{\Gamma(\alpha_{d00} + \alpha_{d01} + \alpha_{d10} + \alpha_{d11})}{\Gamma(\alpha_{d00})\Gamma(\alpha_{d01})\Gamma(\alpha_{d10})\Gamma(\alpha_{d11})} \\
&\quad \times \int p_{d00}^{r_{d00}+\alpha_{d00}} p_{d01}^{r_{d01}+\alpha_{d01}} p_{d10}^{r_{d10}+\alpha_{d10}} p_{d11}^{r_{d11}+\alpha_{d11}} d\boldsymbol{p}_d \\
&= \binom{n_d}{r_{d00}, r_{d01}, r_{d10}, r_{d11}} \frac{\Gamma(\alpha_{d00} + \alpha_{d01} + \alpha_{d10} + \alpha_{d11})}{\Gamma(\alpha_{d00})\Gamma(\alpha_{d01})\Gamma(\alpha_{d10})\Gamma(\alpha_{d11})} \\
&\quad \times \frac{\Gamma(\alpha_{d00} + r_{d00})\Gamma(\alpha_{d01} + r_{d01})\Gamma(\alpha_{d10} + r_{d10})\Gamma(\alpha_{d11} + r_{d11})}{\Gamma(\alpha_{d00} + r_{d00} + \alpha_{d01} + +r_{d01} + \alpha_{d10} + r_{d10} + \alpha_{d11} + r_{d11})}.
\end{aligned} \tag{4}$$

Since the data and prior parameters among cases and controls are assumed to be independent, the joint prior predictive distribution of $\boldsymbol{R}_0$ and $\boldsymbol{R}_1$ is simply the product of the prior predictive distributions of $\boldsymbol{R}_0$ and $\boldsymbol{R}_1$. We consider the following four Bayesian SSD criteria:

(a) **The average posterior variance (APV):** For given $\gamma > 0$, we choose the smallest n satisfying

$$\mathbb{E}\{\text{var}[\beta \mid \boldsymbol{R}_0, \boldsymbol{R}_1]\} \leq \gamma,$$

where $\text{var}[\beta \mid \boldsymbol{R}_0, \boldsymbol{R}_1]$ represent the posterior variance of β given data. Based on the previous representation of β in terms of the difference of G–E log OR in cases and controls, this implies that $\mathbb{E}\{\text{var}[\log(\psi_1) \mid \boldsymbol{R}_1]\} + \mathbb{E}\{\text{var}[\log(\psi_0) \mid \boldsymbol{R}_0]\} \leq \gamma$. We present a closed-form expression for the approximate posterior variance in the Web Appendix.

(b) **The average length criterion (ALC):** Let $L_\alpha(\boldsymbol{R}_0, \boldsymbol{R}_1)$ denote the random length of the 100(1-α)% level highest posterior density (HPD) interval for β. For a given $l > 0$, we find the smallest n satisfying

$$\mathbb{E}\{L_\alpha(\boldsymbol{R}_1, \boldsymbol{R}_0)\} \leq l.$$

(c) **Length probability criterion (LPC):** The ALC ensures that on an *average* one does not obtain large noninformative HPD intervals. However, it does not

Table 3

The estimated bias and MSE corresponding to the different estimation procedures under four simulation settings. The four settings correspond to the following true values for the parameters $(P_G, P_E, OR_{01}, OR_{10}, \psi_0, \psi)$: (*i*) (0.3, 0.5, 1, 1, 1, 1.3); (*ii*) (0.3, 0.5, 1, 1, 1.25, 1.3); (*iii*) (0.36, 0.13, 1.02, 1.06, 0.76, 1.48); *and* (*iv*) (0.36, 0.44, 1.02, 0.87, 0.74, 1.19). *Setting* (*i*) *reflects G–E independence* ($\psi_0 = 1$) *and setting* (*ii*) *reflects moderate but positive G–E dependence* ($\psi_0 = 1.25$). *Whereas settings* (*iii*) *and* (*iv*) *exactly replicate the observed data in MECC study for NAT2* × *RMGF and NAT2* × *SMOKE interaction, where negative association is observed between G and E. CC, CO, and EB refer to the case–control, case-only, and empirical Bayes estimator of* β *while under the FB analysis, we use a fairly noninformative prior on the G–E log OR in cases, namely,* $\log(\psi_1)$, *and vary the strength of the prior on G–E log OR among controls, namely,* $\log(\psi_0)$, *to reflect different levels of uncertainty on gene–environment independence. Results based on* 5000 *simulated datasets under each setting.*

			Simulated 1		Simulated 2		NAT2 × RMGF		NAT2 × SMOKE	
n_1	n_0	Method	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
		CC	0.007	0.200	0.001	0.200	0.023	0.400	0.002	0.181
		CO	0.006	0.098	0.228	0.148	−0.291	0.256	−0.304	0.181
200	200	EB	0.008	0.131	0.101	0.152	−0.112	0.288	−0.126	0.156
		FB 20†	0.013	0.176	0.075	0.167	−0.076	0.226	−0.015	0.146
		FB 80	0.018	0.122	0.087	0.147	−0.101	0.249	−0.098	0.124
		FB 320	0.016	0.104	0.156	0.141	−0.247	0.253	−0.202	0.125
		CC	−0.007	0.076	0.003	0.077	0.004	0.153	0.003	0.071
		CO	0.000	0.037	0.227	0.090	−0.285	0.146	−0.299	0.124
500	500	EB	−0.004	0.049	0.092	0.070	−0.112	0.129	−0.100	0.074
		FB 20	0.013	0.053	0.034	0.075	−0.064	0.121	−0.004	0.065
		FB 80	0.012	0.048	0.060	0.069	−0.111	0.114	−0.050	0.060
		FB 320	0.008	0.040	0.108	0.064	−0.204	0.136	−0.127	0.075
		CC	0.004	0.037	0.000	0.038	0.010	0.073	−0.001	0.035
		CO	0.001	0.019	0.225	0.069	−0.279	0.110	−0.297	0.105
1000	1000	EB	0.002	0.025	0.077	0.040	−0.091	0.072	−0.078	0.042
		FB 20	0.014	0.027	0.016	0.037	−0.027	0.064	−0.004	0.034
		FB 80	0.013	0.024	0.032	0.035	−0.089	0.074	−0.030	0.032
		FB 320	0.011	0.018	0.076	0.034	−0.185	0.107	−0.101	0.036
		CC	0.000	0.019	−0.002	0.019	0.001	0.036	0.001	0.018
		CO	0.001	0.009	0.222	0.058	−0.282	0.096	−0.298	0.098
2000	2000	EB	0.000	0.012	0.055	0.022	−0.077	0.042	−0.049	0.022
		FB 20	0.005	0.019	0.006	0.018	−0.018	0.034	−0.001	0.017
		FB 80	0.005	0.018	0.015	0.018	−0.061	0.038	−0.015	0.017
		FB 320	0.004	0.012	0.042	0.013	−0.141	0.091	−0.061	0.019

†For FB, the case Dirichlet parameter α_1 is fixed as (5, 5, 5, 5) with $s_1 = 20$ and values of s_0 at 20, 80, and 320 correspond to the control Dirichlet parameter α_0 being set at (5, 5, 5, 5), (20, 20, 20, 20), and (80, 80, 80, 80) respectively. Under the above setting, prior variance corresponding to $\log(\psi_1)$ is 0.90 whereas prior variance on $\log(\psi_0)$ under $s_0 = 20$, 80, and 320 are 0.90, 0.20, and 0.05, respectively.

control for variability in the random length values $L_\alpha(\boldsymbol{R}_1, \boldsymbol{R}_0)$. A cautious investigator may not be satisfied with the average assurances provided by the ALC criterion. The LPC criterion provides a more conservative choice where we find the smallest n satisfying

$$\mathbb{P}\{L_\alpha(\boldsymbol{R}_1, \boldsymbol{R}_0) \geq l\} \leq \delta,$$

for fixed l and $\delta \in (0, 1)$. Joseph and Bélisle (1997) derive this criterion as a special case of their proposed worse outcome criterion.

(d) **Criterion for hypothesis testing:** Consider, for example, the problem of testing the hypothesis $H_0 : \beta > 0$ versus $H_a : \beta \leq 0$. In a Bayesian paradigm, one selects one of the two hypotheses, by comparing the posterior probability of the null and the alternative. We say that there is evidence in favor of the null (or alternative) hypothesis if $\pi(H_0 \,|\, \boldsymbol{r}_0, \boldsymbol{r}_1) \geq \rho$ (or $\leq 1 - \rho$) where $\rho \in (0.5, 1)$ is a chosen probability level. A state of indecision is reached when $1 - \rho \leq \pi(H_0 \,|\, \boldsymbol{R}_0, \boldsymbol{R}_1) \leq \rho$, when there is neither sufficient evidence to support H_0, or to refute it. The sample size criterion proposed by Verdinelli (1996) and also used in De Santis et al. (2004) is to choose the smallest n such that the probability of an inconclusive decision stays below a certain small threshold $\delta \in (0, 1)$, i.e.,

$$TC1 \stackrel{\text{def}}{=} \mathbb{P}\{1 - \rho \leq \pi(H_0 \,|\, \boldsymbol{R}_0, \boldsymbol{R}_1) \leq \rho\} \leq \delta.$$

In general, one can test any composite null and alternative hypothesis $H_0 : \beta \in \Phi_0$ versus $H_a : \beta \in \Phi_1 = \Phi_0^c$ by comparing $\pi(H_j \,|\, \boldsymbol{r}_1, \boldsymbol{r}_0) = \int_{\Phi_j} \pi(\beta \,|\, \boldsymbol{r}_1, \boldsymbol{r}_0) d\beta$, $j = 0, 1$, and employ the above SSD criterion. Along with the above testing criterion $TC1$, we also evaluate an analogue of frequentist power criterion that is simply an evaluation of $TC2 \stackrel{\text{def}}{=} \mathbb{P}(\pi(H_a \,|\, \boldsymbol{R}_0, \boldsymbol{R}_1) \geq \rho)$ or the chance of selecting the alternative hypothesis with a pre-specified high posterior probability ρ. A discussion of

various SSD criteria for hypothesis testing and model selection can be found in Weiss (1997) and Wang and Gelfand (2002).

Having introduced the ingredients of the SSD problem for G × E, we now proceed to describe the optimization scheme to allocate cases and controls and to determine the optimal total sample size n with respect to each specific criterion.

4. Computational Algorithm

We consider two approaches toward determining the optimal SSD.

(i) The direct allocation of optimal number of cases and controls and optimal sample size: For each possible sample size n, find the best pair of n_1 and n_0, by directly minimizing/optimizing any of the criteria under consideration, namely, (a)–(d). So, for each candidate sample size n, one has to evaluate the chosen criterion for $(n-1)$ configurations $n_1 = 1, \ldots, n-1$ and $n_0 = n - n_1$ and check whether the desired threshold is met. Then repeat this for all n over a grid and find the minimum $n = n_1 + n_0$ for which the pre-specified threshold levels are met.

(ii) The two-step allocation strategy: Another allocation strategy, akin to the one considered in De Santis et al. (2004) is to follow a two-step procedure where at the first step, for each n, one allocates n_1 and n_0 such that the average posterior uncertainty in $\psi_d,\ d = 0, 1$ is the same in cases and controls. When prior information on cases is sharper than the one in controls, this requirement of equal posterior uncertainty intuitively will lead to allocating more controls than cases and vice versa. Thus, at the first step, one chooses the pair $n_1 = (1, \ldots, n-1), n_0 = n - n_1$, such that,

$$|\,\mathbb{E}(\mathrm{var}(\psi_1 \,|\, \boldsymbol{R}_1)) - \mathbb{E}(\mathrm{var}(\psi_0 \,|\, \boldsymbol{R}_0))\,| \tag{5}$$

is closest to zero. The expression for $\mathrm{var}(\psi_d \,|\, \boldsymbol{r}_d),\ d = 0, 1$, is provided in the Web Appendix. One then evaluates any of the chosen criteria (a)–(d) for each n with this *specific* allocation of n_1 and n_0, and chooses the minimum n for which the threshold conditions are met. This second approach has the advantage that for each candidate n, the best split between cases and controls according to (5) remains the same for each criterion and only the second step of searching for the optimal n needs to be implemented separately for each criterion.

For a given n_1 and n_0, for evaluating the $\mathbb{P}$ and $\mathbb{E}$ with respect to the Multinomial-Dirichlet prior predictive distribution in (4), one first has to select the design prior parameters $\boldsymbol{\alpha}_d,\ d = 0, 1$. In the next section, we show how one can choose these prior parameters based on the observed data in Table 2. Given $\boldsymbol{\alpha}_d$ and n_d one can: (I) generate $\boldsymbol{r}_d$ by following the Gibbs sampling algorithm described in Casella and George (1992) by sequentially generating a large number of observations (10,000, say) from the full conditional distributions $p(\boldsymbol{r}_d \,|\, \boldsymbol{p}_d)$ and $\pi(\boldsymbol{p}_d \,|\, \boldsymbol{r}_d)$, as described in (2). As the length of the sequence increases, the samples from r_d converge stochastically to the marginal distribution $\mathbb{P}(\boldsymbol{R}_d)$ as given in (4). The convergence of the chain was assessed through the diagnostic proposed by Geweke (1992). (II) For each realization of $\boldsymbol{r}_d$, generate 10,000 random samples from the posterior distribution of $(X_d, Y_d, Z_d \,|\, \boldsymbol{r}_d)$, which are standard Beta distribution as described in Result 1. (III) Finally, in each of the above 10,000 generations for a given $\boldsymbol{r}_d$, compute ψ_d as a function of $(X_d, Y_d, Z_d \,|\, \boldsymbol{r}_d),\ d = 0, 1$ as given in (3) and compute β in terms of ψ_d. This process generates 10,000 observations from the posterior distribution of β for a given $(\boldsymbol{r}_1, \boldsymbol{r}_0)$. One can now evaluate the length of the HPD interval or any of the other characteristics of the posterior distribution of β, as described in (a)–(d) of Section 3.

After repeating the above Steps (I)–(III) a large number of times, say 50,000 times, one is able to obtain Monte Carlo estimates of the $\mathbb{P}$ and $\mathbb{E}$ involved in each of the criterion. One can then find the minimum n for which they meet the desired threshold requirement.

Remark 3: M'Lan et al. (2006) indicate that there may be substantial error in the Monte Carlo estimates for the criterion function, and propose a curve fitting technique to the criterion function over a grid of sample sizes instead of direct optimization. We recognize this possibility and use a large number of Monte Carlo samples to provide more stable estimates. The approach of M'Lan et al. (2006) could be adapted as well for more efficient computation. However, unlike M'Lan et al., we simplify our search for optimal design by a somewhat different two-step allocation strategy (De Santis et al., 2004) with a closed-form expression for the posterior variance being used to determine the best split for each given n at the first step. Consequently, the SSD obtained by direct optimization of criteria (a)–(d) may not necessarily reside in the subclass that minimize (5), and thus be different from the ones obtained by the two-step procedure. The use of the two-step strategy reduces our computational burden enormously by limiting the design search space at the second stage to only the optimal (n_0, n_1) that are identified at the first stage by minimizing (5) for each given n. This makes repeated evaluation of the SSD criteria at the second step feasible and one can take the average of the replicated values to reduce Monte Carlo errors further.

5. Illustrative Example

For a general discussion regarding the dual issue of choosing design priors versus analysis priors, we refer the reader to (Wang and Gelfand, 2002; Spiegelhalter, Abrams, and Myles, 2004; De Santis, 2006). For specifying the prior hyperparameters based on current MECC study data, we used the same technique applied in Bernardo and Smith (1994) and De Santis et al. (2004). Based on observed data given in Section 2, the prior estimates of $\boldsymbol{\alpha}_d$ were calculated by using the following equation:

$$\frac{\alpha_{dge}}{s_d} = \frac{r_{dge}}{n_d} (d = 0, 1; g = 0, 1; e = 0, 1),$$
$$\frac{s_0}{s_1} = \frac{n_0}{n_1}, \qquad \sum_{d=0}^{1}\sum_{g=0}^{1}\sum_{e=0}^{1} \alpha_{dge} = s. \tag{6}$$

The obtained design prior parameters are displayed in Web Table 4 for RMGF and SMOKE data. The variation in prior

Table 4

The LPC, APC, and APV criterion values corresponding to estimation of NAT2 × RMGF and NAT2 × SMOKE interaction for certain representative values of n. Optimal allocation of cases and controls for each candidate n is as determined in Web Table 5. *For LPC and ALC, we consider a* 95% *HPD credible interval. The Dirichlet prior parameters for each strength are chosen as given in Web Table* 4, *based on data from MECC study. We evaluated the LPC criterion by choosing the threshold length of the* 95% *HPD credible interval as* $l = 0.8$ *for illustration purposes. An investigator can find optimal sample sizes by selecting the smallest n for which the threshold that is desired for his/her criterion is first met or refine the search in that interval. For example, the sample sizes in bold face denote approximately optimal sample sizes if ALC was desired to be approximately* 0.8 *or lower.*

Prior strength	NAT2 × RMGF						NAT2 × SMOKE					
	n	n_1	n_0	LPC	ALC	APV	n	n_1	n_0	LPC	ALC	APV
	500	202	298	1.000	1.870	0.232	500	256	244	1.000	1.259	0.104
	1000	412	588	1.000	1.428	0.136	1000	504	496	1.000	0.961	0.061
	2000	828	1172	0.999	1.060	0.075	**1500**	754	746	0.606	**0.808**	0.043
200	3000	1262	1738	0.795	0.878	0.052	1700	848	852	0.061	0.764	0.038
	3500	1514	1986	0.521	0.816	0.045	2000	998	1002	0.010	0.711	0.033
	3700	1594	2106	0.416	**0.795**	0.042	2500	1250	1250	<0.001	0.642	0.027
	4000	1728	2272	0.258	0.760	0.039	3000	1500	1500	<0.001	0.590	0.023
	500	184	316	1.000	1.502	0.149	500	268	232	1.000	1.041	0.071
	1000	402	598	1.000	1.229	0.100	1000	516	484	1.000	0.851	0.048
	1500	620	880	1.000	1.062	0.075	**1300**	664	636	0.063	**0.777**	0.040
500	2000	836	1164	1.000	0.953	0.060	1500	762	738	0.032	0.737	0.036
	2500	1054	1446	0.933	0.870	0.050	2000	1012	988	0.009	0.659	0.029
	3000	1266	1734	0.492	**0.803**	0.043	2500	1260	1240	<0.001	0.602	0.024
	3500	1480	2020	0.152	0.754	0.038	3000	1506	1494	<0.001	0.557	0.023
	3800	1606	2194	0.045	0.726	0.035	3500	1746	1754	<0.001	0.534	0.021
	4000	1704	2296	0.023	0.709	0.033	4000	2000	2000	<0.001	0.531	0.020
	500	150	350	1.000	1.214	0.097	500	288	212	1.000	0.847	0.047
	1000	370	630	1.000	1.051	0.073	**700**	386	314	0.357	**0.796**	0.042
	1500	586	914	1.000	0.940	0.058	780	421	349	0.051	0.780	0.040
1000	2000	806	1194	0.977	0.859	0.049	1000	536	464	0.013	0.734	0.035
	2500	1018	1482	0.423	**0.795**	0.042	1500	782	718	<0.001	0.657	0.028
	2900	1198	1702	0.060	0.753	0.038	2000	1030	970	<0.001	0.599	0.024
	3000	1240	1760	0.032	0.743	0.036	2500	1278	1222	<0.001	0.555	0.020
	3500	1456	2044	0.002	0.701	0.032	3000	1526	1474	<0.001	0.519	0.018
	4000	1680	2320	<0.001	0.664	0.029	4000	2000	2000	<0.001	0.432	0.014

strength controls the prior variance, with the prior mean for β and all other parameters being fixed at the corresponding MLE obtained from the current MECC database. The details corresponding to each setting are provided in Web Table 4.

Based on the prior parameters as chosen above, the optimal pairs can be selected via either method (i) or (ii) previously explained. We present the best pairs for some candidate $n = n_1 + n_0$ obtained in the first step of the two-stage approach in Web Table 5 by equating the expected posterior uncertainty in cases and controls for both SMOKE and RMGF data, respectively. One can notice that for each n the number of cases and controls needed is almost even for NAT2 × SMOKE interaction, whereas more controls are needed for the NAT2 × RMGF interaction. In fact, the large sample asymptotic variances corresponding to the MLEs of $\log(\psi_0)$ and $\log(\psi_1)$ are 0.0202 and 0.0175 for RMGF and 0.009 and 0.0098 for SMOKE data, intuitively justifying the reason for this type of allocation based on the information contained in the design prior. Note that the effect of prior on the optimal allocation rule is more pronounced for smaller n than larger n.

To get a sense of the nature of the different criteria, we report the criterion values for different candidate n in Table 4, with the optimal pairs as determined by Web Table 5. The three criteria LPC, ALC, and APV focus on estimation accuracy. For the length-based criteria, we consider 95% HPD intervals. One can notice as expected that the LPC criterion provides a larger sample size than the ALC criterion as the former is controlling the probability distribution of the random interval lengths whereas the latter is simply controlling for the mean length. With increasing prior strength the value of each criterion decreases and the required n needed to attain a certain threshold becomes smaller. Comparing the results for RMGF with the results for SMOKE one can notice that everything else remaining same, much larger sample sizes are needed for RMGF, to reach the same level of precision when compared to SMOKE. Figure 2 presents a graphical display of the ALC and LPC criterion with $l = 0.8$, plotted against each candidate n that also illustrates the same features of the SSD summarized above.

Web Table 6 presents the optimal pair (n_1, n_0) obtained by the direct approach (i) by evaluating the ALC criterion instead of the two-stage allocation approach. Though there

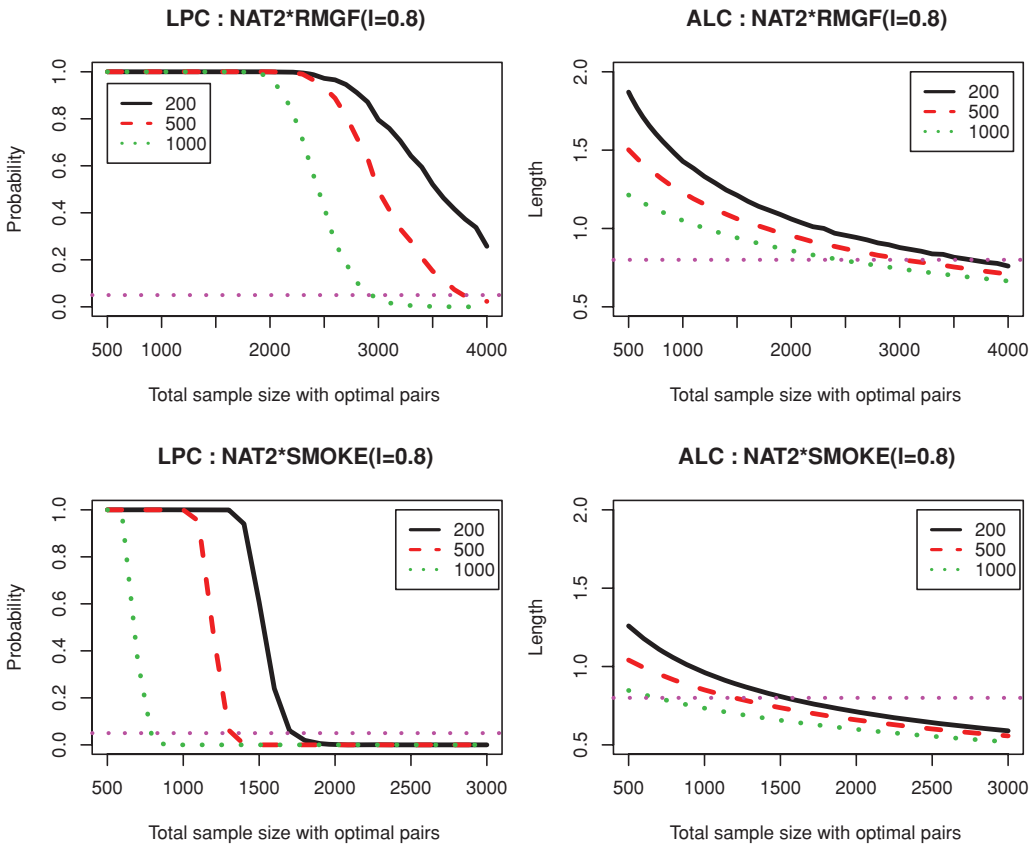

Figure 2. A graphical display of LPC and ALC criterion, plotted against each candidate n for estimating NAT2 × RMGF and NAT2 × SMOKE interaction log OR parameter. Optimal allocation of cases and controls for each candidate n is as determined in Web Table 5 by using equation (5). In each case, we consider a 95% HPD credible interval. The Dirichlet prior parameters for each strength are chosen as given in Web Table 2, based on data from MECC study. We chose the threshold values as $l = 0.8$ and $\delta = 0.05$, where the horizontal lines on each graph are drawn. The n at which the graph first exceeds the horizontal line is the desired optimal sample size. Corresponding numerical results are collected in Table 4. This figure appears in color in the electronic version of this article.

are certain numerical differences, the main comparative features between the two datasets remain the same. The trend in differences in number of cases and controls is less obvious in Web Table 6 than the two-stage approach (Web Table 5) as the first-step criterion of equating posterior uncertainty in the two-step approach is not imposed in the direct optimization approach.

We would like to point out that the criteria considered in Table 4 only focus on variance but not the aspect of detecting given departures from a null hypothesis or the effect size. In Table 5, we present the results on the two testing criteria we discussed in (d) for testing $H_0 : \beta \leq 0$ vs $H_a : \beta > 0$. Notice that with the test criterion TC1 that minimizes the probability of reaching the indifference zone, one can actually need more sample size with a more informative prior with higher strength (for NAT2 × SMOKE data in Table 5, using TC1 ≤ 0.05 as the desired criterion, $s = 200 : \ n_1 = 1500, \ n_0 = 1500$; with $s = 500 : \ n_1 = 2761, \ n_0 = 2739$). This may seem quite surprising, but has also been observed in De Santis et al. (2004). The explanation lies in the fact that for the precision-based criteria, larger number of observations are needed in absence of strong prior information to reach the target precision. However, in the testing criterion the location and dispersion of the posterior both play a role in determining the sample size. The sample observations also help to locate the posterior density away from the prior mean, and often in absence of strong prior information, few observations are enough to shift the posterior density to conclude in favor of one of the hypotheses. We also notice that for $TC1$, both RMGF and SMOKE require approximately similar sample sizes for weaker prior strength, RMGF needing much smaller sample sizes for higher prior strength. This is also expected as in our design prior, RMGF shows stronger departure from the null than the prior chosen for SMOKE, reflecting the findings of the current MECC study. In evaluating the analogue of the frequentist power criterion $TC2$, we evaluate the probability of concluding H_a with posterior odds larger than 3:1 in favor of H_a. This probability remains consistently higher for NAT2 × RMGF interaction, than NAT2 × SMOKE interaction in Table 5.

In Table 6, we evaluate the Bayesian SSD criteria under a CC design that will typically be used to ensure 80% power with a type-1 error level of 5% in a classical frequentist setting for testing $H_0 : \beta = 0$ against $H_a : \beta \neq 0$. We consider four

Table 5

*The two test criterion values (TC*1 *and TC*2*) corresponding to estimation of NAT*2 × *RMGF and NAT*2 × *SMOKE interaction. Optimal allocation of cases and controls for each candidate n is as determined in Web Table* 5*. The Dirichlet prior parameters for each strength are chosen as given in Web Table* 4*, based on data from MECC study. In both cases, we consider* $H_0 : \beta \leq 0$ *versus* $H_a : \beta > 0$*. The criterion* $\text{TC1} = \mathbb{P}\{0.4 \leq \pi(H_a \mid \boldsymbol{R}_0, \boldsymbol{R}_1) \leq 0.6\}$ *evaluates the probability of falling in an indecisive zone, whereas the criterion* $\text{TC2} = \mathbb{P}\{\pi(H_a \mid \boldsymbol{R}_0, \boldsymbol{R}_1) \geq 0.75\}$ *evaluates the probability of concluding* H_a *with posterior odds larger than* 3:1*. One would like to determine sample sizes with small values of TC*1 *(say* 0.05*, for example) and higher values of TC*2*, the posterior probability of detecting a positive interaction effect.*

Prior strength	NAT2 × RMGF					NAT2 × SMOKE				
	n	n_1	n_0	TC1	TC2	n	n_1	n_0	TC1	TC2
200	500	202	298	0.105	0.568	500	256	244	0.123	0.468
	1000	412	588	0.079	0.601	1000	504	496	0.089	0.516
	1500	630	870	0.066	0.603	1500	754	746	0.073	0.522
	2000	828	1172	0.048	0.620	2000	998	1002	0.064	0.527
	2500	1046	1454	0.052	0.628	2500	1250	1250	0.054	0.550
	3000	1262	1738	0.049	0.636	3000	1500	1500	0.050	0.554
	4000	1683	2317	0.040	0.633	3500	1750	1750	0.048	0.547
	5000	2103	2897	0.040	0.635	4000	2000	2000	0.044	0.556
	6000	2524	3476	0.035	0.638	5000	2500	2500	0.036	0.577
	7000	2945	4055	0.029	0.655	6000	3000	3000	0.040	0.571
	8000	3365	4635	0.030	0.656	7000	3500	3500	0.032	0.587
	10000	4207	5793	0.026	0.665	10000	5000	5000	0.030	0.595
500	500	184	316	0.116	0.649	500	268	232	0.158	0.504
	1000	402	598	0.090	0.676	1000	516	484	0.127	0.541
	1500	620	880	0.078	0.697	1500	762	738	0.096	0.564
	2000	836	1164	0.069	0.695	2000	1012	988	0.084	0.583
	2400	1010	1390	0.065	0.700	2500	1260	1240	0.081	0.580
	3000	1266	1734	0.064	0.703	3000	1506	1494	0.074	0.601
	4000	1688	2312	0.053	0.710	4000	2008	1992	0.063	0.606
	5000	2110	2890	0.045	0.727	5000	2510	2490	0.064	0.615
	6000	2532	3468	0.043	0.718	5500	2761	2739	0.051	0.615
	7000	2954	4046	0.046	0.728	7700	3850	3850	0.048	0.624
	8000	3376	4624	0.037	0.727	8000	4000	4000	0.045	0.628
	9000	3798	5202	0.036	0.729	9000	4500	4500	0.041	0.633
	10000	4220	5780	0.034	0.738	10000	5000	5000	0.041	0.639

parameter settings identical to Table 3. Table 6 gives one a sense of the thresholds for the Bayesian SSD criteria that are comparable with corresponding frequentist power criteria. There are certain numerical differences in the Bayesian criteria *TC*2 and ALC from simulation setting (i) (independence) to (ii) (dependence) under increasing strength in the design prior. In general, the criteria values remain fairly robust. Since the Bayesian and frequentist testing paradigms are very different, it is hard to directly translate the two sets of operating characteristics. We evaluate the frequentist type-I error and power corresponding to the Bayesian decision rule employed in TC2 where the null hypotheses is rejected if $\pi(H_a \mid \boldsymbol{r}_0, \boldsymbol{r}_1) \geq 0.75$. We include performances of the Wald tests based on the CO and EB estimators under this design. The unacceptable performance of the CO estimator under departures from *G–E* independence illustrates why the assumption should not be taken for granted at the design stage if there is reasonable doubt and the need for allowing uncertainty around this assumption. Especially under negative *G–E* dependence (as also shown in Figure 1 of Li and Conti, 2009), CO even loses its power advantages along with unbelievably high type-I error rates. The EB procedure strikes a compromise and is more robust across *G–E* association scenarios.

In terms of setting specific values for the different Bayesian SSD criteria, there are no established benchmarks such as 80% power or 5% type-I error in the classical setting, and largely depend on the investigator and the study under consideration. In Table 7, we specify certain ad hoc thresholds for the ALC, LPC, and TC2 criteria and study the changes in the SSD across prior choices for different simulation settings. One can notice the changes in the SSD requirements when estimation accuracy is the goal (ALC, LPC) in contrast to detecting an effect size (TC2) and also the reduced sample sizes with increase in prior strength. To reiterate that the requirements in Bayesian perspective could be quite different from that in a frequentist testing situation, we provide the frequentist power of (CC, CO, EB) methods under the Bayesian design.

Of possible interest is the design situation when one chooses all available number of cases and wants to select the optimal case:control ratio in terms of a given criterion. Web Table 7 provides evaluation of the ALC criterion at certain designated number of cases and varying case:control ratios of 1:1–1:3 from which an investigator can determine the optimal ratio for a

Table 6

Comparison of Bayes SSD criteria under frequentist design: The four Bayes SSD criteria estimates under two simulated data configurations, and data corresponding to inference regarding NAT2 × RMGF and NAT2 × SMOKE interaction are presented. The sample sizes are pre-determined to achieve a power of 0.8, *with level of significance* 5% *to detect a significant interaction* β *under a classical case–control design. The two simulated datasets with* 2000 *cases and* 2000 *controls were created to yield the value of* $(P_G, P_E, OR_{01}, OR_{10}, \psi_0, \psi)$ *as* (*i*) (0.3, 0.5, 1, 1, 1, 1.3) *and* (*ii*) (0.3, 0.5, 1, 1, 1.25, 1.3), *reflecting G–E independence and G–E dependence, respectively. For NAT2 × RMGF and NAT2 × SMOKE interaction, the parameter guesses are fixed at the point estimates obtained in the MECC study, namely, at* (0.36, 0.13, 1.02, 1.06, 0.76, 1.48) *and* (0.36, 0.44, 1.02, 0.87, 0.74, 1.19,) *respectively. The criterion considered for estimation are ALC and APV. For Bayesian hypotheses testing, we consider the hypotheses* $H_0 : \beta \leq 0$ *versus* $H_a : \beta > 0$. *The testing criterion* $TC1 = \mathbb{P}\{0.4 \leq \pi(H_a \mid \boldsymbol{R}_0, \boldsymbol{R}_1) \leq 0.6\}$ *evaluates the probability of falling in an indecisive zone, whereas the criterion* $TC2 = \mathbb{P}\{\pi(H_a \mid \boldsymbol{R}_0, \boldsymbol{R}_1) \geq 0.75\}$ *evaluates the probability of concluding* H_a *with posterior odds larger than* 3:1. *In the last two columns, we also add the performance of Wald tests based on the case-only estimator and the empirical Bayes estimator under each simulation setting, with Power* (*Type-I error in parentheses*) *under the given sample sizes.*

Data	Prior strength	$n_1(= n_0)$	TC1	TC2(Power*, Type 1 error*)	ALC	APV	Power (Type 1 error) CO	Power (Type 1 error) EB
Simulated 1	200	4300	0.030	0.659(0.965,0.242)	0.357	0.009	0.981 (0.054)	0.918 (0.042)
	500		0.038	0.752(0.974,0.241)	0.336	0.007		
	1000		0.039	0.833(0.987,0.226)	0.316	0.006		
Simulated 2	200	4300	0.030	0.641(0.963,0.250)	0.360	0.009	1.000 (0.911)	0.862 (0.066)
	500		0.038	0.721(0.974,0.238)	0.359	0.008		
	1000		0.040	0.796(0.986,0.215)	0.338	0.007		
RMGF	200	3600	0.032	0.662(0.989,0.251)	0.582	0.022	0.268 (0.761)	0.623 (0.083)
	500		0.034	0.733(0.995,0.234)	0.544	0.019		
	1000		0.036	0.827(0.998,0.212)	0.521	0.018		
SMOKE	200	8900	0.020	0.605(0.984,0.252)	0.249	0.004	0.811 (1.000)	0.705 (0.051)
	500		0.029	0.663(0.989,0.243)	0.244	0.004		
	1000		0.034	0.701(0.994,0.242)	0.240	0.004		

* The power and type-I error corresponding to $TC2$ are calculated based on the fraction of time the Bayesian decision rule $\pi(H_a \mid \boldsymbol{r}_0, \boldsymbol{r}_1) \geq 0.75$ rejects the null hypothesis $H_0 : \beta < 0$, with $\boldsymbol{r}_0$ and $\boldsymbol{r}_1$ being generated from the multinomial distributions with the four parameter settings described above, under the alternative ($\psi = 1.3$) and the null ($\psi = 1$), respectively.

chosen threshold. M'Lan et al. (2006) point out that with a very diffuse design prior, the sample sizes provided by the ALC criteria are unstable, thus we have refrained from using a very diffuse Dirichlet prior as a design prior in the current study.

In conclusion, the results in terms of optimal allocation, as well as controlling the different aspects of the posterior illustrate the appealing data-adaptive feature of Bayesian SSD, by including prior information as well as uncertainty from an existing study into designing a new study. Also, by varying the prior strength one can calibrate one's faith in the findings of the current study and control the effective sample size. Bayesian computation becomes extremely fast due to inexpensive generation of the posterior samples. Software codes written in `R` for the Bayesian analysis and for determining the optimal SSD are available at `http://www.sph.umich.edu/bhramar/public_html/research`.

6. Discussion

The current article presents a FB approach to analyze and design studies of gene–environment interaction. The FB analysis of interactions presented in the article is extremely straightforward, but have not previously been indicated in the literature. The article illustrates how to incorporate prior uncertainties around the assumption of gene–environment independence in an FB framework as an alternative to previously proposed EB strategy. The FB formulation allows us to explore the SSD problem for G × E studies in a proper Bayesian framework. We illustrate through use of data from an ongoing CC study how to formulate a design prior in order to plan a future study. An ensemble of Bayesian SSD criteria are considered and evaluated. The methods are compared with other existing design and analytic choices, thus making the article the first comprehensive work on proper full Bayesian analysis and design for G × E problems after the initial exposition in M'Lan et al. (2006). In the absence of existing data or historical information, choosing a design prior is a difficult task, just like fixing effect sizes agnostically in a traditional power calculation is, and the article does not propose a remedy for that situation. In such situations, the newly proposed robust Bayesian SSD criteria that protect against the choice of a single base prior (Brutti, DeSantis, and Gubbiotti, 2008) may be gainfully employed.

The methods can be directly applied toward studies of gene–gene interactions where similar issues regarding gene–gene independence may arise. The model could be extended to accommodate a general categorical G, such as genotype data for single nucleotide polymorphisms (SNP) that are typically coded as 0, 1, or 2 by counting the number of variant alleles carried by an individual. A limitation of the proposed analysis is that the methods are restricted to binary/categorical G and E. However, it is often the case that approximate

Table 7

The variation in sample sizes under three Bayesian SSD criteria: (i) $ALC \leq 0.6$, (ii) $LPC \leq 0.05$, and (iii) $TC2 = \mathbb{P}\{\pi(H_a \mid \boldsymbol{R}_0, \boldsymbol{R}_1) \geq 0.75\} \geq 0.7$. We evaluated the LPC criterion by choosing the threshold length of the 95% HPD credible interval as $l = 0.6$. CC, CO, and EB refer to the case–control, case-only, and empirical Bayes estimator of β and we provide power of these procedures under the chosen Bayesian design. The two simulated datasets were created to yield the value of $(P_G, P_E, OR_{01}, OR_{10}, \psi_0, \psi)$ as (i) (0.3, 0.5, 1, 1, 1, 1.3) and (ii) (0.3, 0.5, 1, 1, 1.25, 1.3), reflecting G–E independence and G–E dependence, respectively. For NAT2 × RMGF and NAT2 × SMOKE interaction, the parameter guesses are fixed at the point estimates obtained in the MECC study, namely, at (0.36, 0.13, 1.02, 1.06, 0.76, 1.48) *and* (0.36, 0.44, 1.02, 0.87, 0.74, 1.19), *respectively.*

				Power		
Data	Prior strength	n_1	n_0	CC	CO†	EB
ALC* ≤ 0.6						
NAT2 × RMGF	200	1935	2565	0.601	0.161	0.404
	500	1677	2223	0.536	0.132	0.335
NAT2 × SMOKE	200	1050	1050	0.155	0.151	0.086
	500	900	900	0.141	0.148	0.076
Simulated Data 1	200	1316	1484	0.266	0.468	0.373
	500	1128	1272	0.234	0.419	0.336
Simulated Data 2	200	1610	1890	0.321	0.989	0.481
	500	1426	1674	0.290	0.977	0.457
LPC* ≤ 0.05				CC	CO	EB
NAT2 × RMGF	200	4128	5472	0.901	0.270	0.788
	500	3225	4275	0.812	0.212	0.640
NAT2 × SMOKE	200	1750	1750	0.229	0.229	0.122
	500	1400	1400	0.201	0.183	0.106
Simulated Data 1	200	2162	2438	0.415	0.682	0.573
	500	1739	1961	0.363	0.596	0.492
Simulated Data 2	200	2070	2430	0.401	0.997	0.530
	500	1656	1944	0.335	0.987	0.479
TC2** ≥ 0.7				CC	CO	EB
NAT2 × RMGF	200	5160	6840	0.953	0.316	0.878
	500	1032	1368	0.346	0.096	0.201
NAT2 × SMOKE	200	12500	12500	0.919	0.907	0.875
	500	11000	11000	0.883	0.862	0.825
Simulated 1	200	5405	6095	0.798	0.971	0.910
	500	2068	2332	0.400	0.661	0.547
Simulated 2	200	5750	6750	0.821	1.000	0.871
	500	2760	3240	0.506	1.000	0.622

*ALC and LPC focus on estimation whereas ** TC2 focuses on detecting departures from a null hypothesis.
†CO has elevated type-I errors except for Simulated 1 scenario ($\psi_0 \equiv 1$), exact type-I error values not presented here.

sample size and power calculations are carried out even for continuous exposures under such stratified setting in many study proposals. A hierarchical Bayesian analysis using the retrospective likelihood framework of Chatterjee and Carroll (2005) under a general regression set-up is possible, but requires use of substantively greater computational time (Mukherjee et al., 2007) and a computationally more challenging SSD problem.

Though the article captures important features of a G × E study design, it focuses on "one at a time" analysis of a single G × E parameter. In the modern era of genome-wide association studies, for testing thousands of genes and their interactions simultaneously, an appropriate utility function that reflects the issue of error due to multiple testing should be considered in a proper Bayesian manner, a topic that remains of future research interest.

7. Supplementary Materials

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

Acknowledgements

The research of Bhramar Mukherjee was partially supported by NSF DMS 07-06935 and NIH grant R03 CA130045-01. The Molecular Epidemiology of Colorectal Cancer Study is supported via NIH grant R01 CA81488. The research of Nilanjan Chatterjee was supported by a Gene-Environment Initiative (GEI) grant from the National Heart Lung and Blood Institute (R01 HL091172-01) and by the Intramural research program of the National Cancer Institute. Malay Ghosh's research was supported in part by NSF grant SES-063426.

References

Adcock, C. J. (1997). Sample size determination: A review. *The Statistician* **46,** 261–283.

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.

Barrett, J. H., Smith, G., Waxman, R., Gooderham, N., Lightfoot, T., Garner, R. C., Augustsson, K., Wolf, C. R., Bishop, D. T., Forman, D., and The Colorectal Cancer Study Group. (2003). Investigation of interaction between N-acetyltransferase 2 and heterocyclic amines as potential risk factors for colorectalcancer. *Carcinogenesis* **24,** 275–282.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

Bernardo, J. M. (1997). Statistical inference as a decision problem: The choice of sample size. *The Statistician* **46,** 151–153.

Brutti, P., DeSantis, F., and Gubbiotti, S. (2008). Robust Bayesian sample size determination in clinical trials. *Statistics in Medicine* **27,** 2290–2306.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician* **46**(3), 167–174.

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

De Santis, F. (2006). Sample size determination for robust Bayesian anlaysis. *Journal of the American Statistical Association* **101,** 278–291.

De Santis, F. (2007). Using historical data for Bayesian sample size determination. *Journal of the Royal Statistical Society, Series A* **170**(1), 95–113.

De Santis, F., Pacifico, M. P., and Sambucini, V. (2004). Optimal predictive sample size for case-control studies. *Applied Statistics* **53**(1), 427–441.

Foppa, I. and Spiegelman, D. (1997). Power and sample size calculations for case-control studies of gene-environment interactions with a polytomous exposure variable. *American Journal of Epidemiology* **146,** 596–604.

Garcia-Closas, M. and Lubin, J. (1999). Power and sample size calculations in case-control studies of gene-environment interactions: Comments on different approaches. *American Journal of Epidemiology* **149,** 689–692.

Gauderman, W. J. (2002). Sample size requirements for matched case-control studies of gene-environment interaction. *Statistics in Medicine* **21,** 35–50.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4*, J. M. Bernado, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds), 169–188. Oxford, UK: Clarendon Press.

Ghosh, J. K., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis*. New York: Springer.

Goldstein, A., Falk, R., Korczak, J., and Lubin, J. (1997). Detecting gene-environment interactions using a case-control design. *Genetic Epidemiology* **14,** 1085–1089.

Gustafson, P., Le, N., and Vallee, M. (2002). A Bayesian approach to case-control studies with errors in covariables. *Biostatistics* **3,** 229–243.

Hein, R., Beckmann, L., and Chang-Claude, J. (2008). Sample size requirements for indirect association studies of gene-environment interactions (G × E). *Genetic Epidemiology* **32,** 235–245.

Hwang, S., Beaty, T., Liang, K., Coresh, J., and Khoury, M. (1994). Minimum sample size estimation to detect gene-environment interaction in case-control designs. *American Journal of Epidemiology* **140,** 1029–1037.

Joseph, L. and Bélisle, P. (1997). Bayesian sample size determination for normal means and differences between normal means. *The Statistician* **46,** 209–226.

Joseph, L., du Berger, R., and Bélisle, P. (1997). Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics in Medicine* **16,** 769–781.

Joseph, L. and Wolfson, D. B. (1997). Interval-based versus decision theoretic criteria for the choice of sample size. *The Statistician* **46,** 145–149.

Joseph, L., Wolfson, D., and du Berger, R. (1995). Sample size calculations for binomial proportions via highest posterior density intervals. *The Statistician* **44,** 143–154.

Kraft, P., Yen, Y. C., Stram, D. O, Morrison, J., and Gauderman, W. J. (2007). Exploiting gene-environmental interaction to detect genetic associations. *Human Heredity* **63,** 111–119.

Latorre, G. (1982). The exact posterior distribution of the cross-ratio of a 2 × 2 contingency table. *Journal of Statistical Computation and Simulation* **16,** 19–24.

Latorre, G. (1984). Bayesian inference in 2 × 2 and 2 × 2 × 2 contingency tables. *Metron* **42,** 169–184.

Lecoutre, B. and Charron, C. (2000). Bayesian procedures for prediction analysis of implication hypothesis in 2 × 2 contingency tables. *Journal of Educational and Behavioral Statistics* **25,** 185–201.

Li, D. and Conti, D. V. (2009). Detecting gene-environment interactions using a combined case-only and case-control approach. *American Journal of Epidemiology* **169,** 497–504.

Lindley, D. V. (1964). The Bayesian analysis of contingency tables. *The Annals of Mathematical Statistics* **35,** 1622–1643.

Lindley, D. V. (1997). The choice of sample size. *The Statistician* **46,** 129–138.

Luan, J. A., Wong, M. Y., and Day, N. E. (2001). Sample size determination for studies of gene-environment interaction. *International Journal of Epidemiology* **30,** 1049–1054.

M'Lan, C. E., Lawrence, L., and Wolfson, B. D. (2006). Bayesian sample size determination for case-control studies. *Journal of the American Statistical Association* **101**(474), 760–772.

Mukherjee, B. and Chatterjee, N. (2008). Exploiting gene-environment independence for analysis of case-control studies: An empirical-Bayes type shrinkage estimator to trade off between bias and efficiency. *Biometrics* **64,** 685–694.

Mukherjee, B., Zhang, L., Ghosh, M., and Sinha, S. (2007). Semiparametric Bayesian analysis of case-control data under conditional gene-environment independence. *Biometrics* **63,** 834–844.

Mukherjee, B., Ahn, J., Gruber, S. B., Rennert, G., Moreno, V., and Chatterjee, N. (2008). Testing gene-environment interaction from case-control data: A novel study of Type-1 error, power and designs. *Genetic Epidemiology* **32,** 615–626.

Müller, P. and Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika* **84,** 523–537.

Pham-Gia, T. (1997). On Bayesian analysis, Bayesian decision theory and the sample size problem. *The Statistician* **46,** 139–144.

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.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Boston: Harvard University Press.

Roberts-Thomson, I. C., Ryan, P., Khoo, K. K., Hart, W. J., McMichael, A. J., and Butler, R. N. (1996). Diet, acetylator phenotype and risk of colorectal neoplasia. *Lancet* **347,** 1372–1374.

Sinha, S., Mukherjee, B., and Ghosh, M. (2004). Bayesian semiparametric modeling for matched case-control studies with multiple disease states. *Biometrics* **60,** 41–49.

Sinha, S., Mukherjee, B., Ghosh, M., Mallick, B. K., and Carroll, R. J. (2005). Bayesian semiparametric analysis of matched case-control studies with missing exposure. *Journal of the American Statistical Association* **100,** 591–601.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health Care Evaluation*. Chichester: Wiley.

Verdinelli, I. (1996). Bayesian design of experiments for the linear model. PhD Thesis. Department of Statistics, Carnegie Mellon university, Pittsburgh.

Walker, S. G. (2003). How many samples? A Bayesian nonparametric approach. *The Statistician* **52,** 475–482.

Wang, F. and Gelfand, A. E. (2002). A simuldation-based approach to Bayesian sample size determination for performance under a given model and separating models. *Statistical Sciences* **17,** 193–208.

Weiss, R. (1997). Bayesian sample size calculation for hypothesis testing. *The Statistician* **46,** 185–191.

Yang, Q., Khoury, M. J., and Flanders, W. D. (1997). Sample size requirements in case-only designs to detect gene-environment interaction. *American Journal of Epidemiology* **146,** 713–720.

Received August 2008. *Revised July* 2009.
Accepted August 2009.