Volume 70, Issue 3, 2026
Journal of Scientific Research
of
The Banaras Hindu University

Understanding the Epidemiology of Genes and Environmental Factors in the Occurrence of any Disease: A Bayesian Study

Akanksha Gupta[1], Swarnalata Tarai*[2], and S. K. Upadhyay[3]

[1]Department of Statistics, Banaras Hindu University, akankshagupta@bhu.ac.in
*[2]Department of Statistics, Banaras Hindu University, swarnalatatarai5@gmail.com
[3]Department of Statistics, Banaras Hindu University, skupadhyay@gmail.com

***Abstract*—The article focuses on complete Bayes analysis of an epidemiological problem assuming a multinomial-Dirichlet modelling assumption and attempts to provide inferences on various parameters including the parameter reflecting gene-environment interaction. It considers the analysis of a $2 \times 2 \times K$ case-control design where all the K factors are representing two levels each. Besides drawing the main inferences on odds ratio and interaction parameters, attempts are also made to provide appropriate choices for the hyperparameters of the assumed Dirichlet prior, an issue that often bothers the Bayesian analysts.**

***Index Terms*—Bayes analysis, Case-control, Dirichlet, Multinomial, Odds ratio.**

I. Introduction

Disease in human population occurs sometimes due to one's physical structure inherited by birth and sometimes due to outer environmental components that become part of individual's everyday life. In genetic epidemiology it is frequently observed that disease clusters in families but an individual family member may or may not inherit the disease. Instead, the individual may inherit sensitivity to certain environmental factors broadly defined to include infectious, chemical, physical, nutritional and behavioural factors. An individual may be differently affected by exposure to the same environment. Subtle differences in genetic factors also cause people to respond differently to the same environmental exposure. This explains why some individuals have a fairly low risk of developing a disease as a result of an environmental insult, while others are much more vulnerable.

The study of the effect of this interaction may not be easy because of the complications involved at various stages and also because of the specific requirement of data (see, for example, [15]). This is perhaps the reason that researchers have normally focused on the analysis assuming gene and environmental components to be independent.

[15] pointed out that it may be often realistic to assume genetic susceptibility and environmental exposure to be independent and, as such, there may be scope of case-only design for drawing the desired inferences. Although the independence assumption may appear reasonable, there may not be sufficient empirical data to support such an assumption. As a matter of fact, case-control study design is strongly recommended and used especially when there is a possibility of departure from the independence. A review of the literature suggests that we have numerous studies related to case-control study designs on genetic susceptibility and the environment exposure especially with reference to rare diseases. These studies are performed using both classical and Bayesian tools (see, for example, [15],[16], etc.). Whatever is the study design under consideration, it is obvious that data are always available in the form of counts and the same can be represented as binomial or multinomial cell frequencies with unknown probabilities of falling in a particular cell.

Among a few important classical references on the impact of genetic susceptibility and environmental exposure, one can refer to [14], who proposed a simple case-control analysis for drawing inferences on odds ratio in the study of ovarian cancer patients. [11] is another important classical reference where the authors provided an easy-to-implement approach for analyzing case-control and case-only study designs under the assumption of gene-environment independence and Hardy-Weinberg equilibrium assumption. The authors called the approach as 'conditional logistic regression with counterfactuals' and successfully established that the approach has several important advantages over the usual approaches.

On Bayesian front the work appears to be relatively meagre although the new researches are going on very rapidly. Among the important Bayesian references, [2] considered data on cases and controls and assumed beta-binomial modelling assumption to draw the desired inferences on odds ratio. [1]

DOI:10.37398/JSR.2026.700309

provided a multivariate extension of the work done by [2]. They used multinomial-Dirichlet modelling assumption and studied instead the effect of covariates in the bile acids of gallbladder diseases. An important problem with the Bayesian analysis given in the above references is the exact assessment of prior hyperparameters. The authors mostly used some logically adopted *ad hoc* mechanisms based on sensitivity analysis to recommend for the appropriate choices of prior hyperparameters. [15] is an important reference in this area where the authors proposed an EB approach to study the interaction between genetic and environmental components.

[16] provided a full Bayesian analysis using multinomial-Dirichlet modelling assumption for the same binary structured case-control scenario with both genetic and environmental components. The authors finally concentrated on obtaining the estimated posterior distribution of multiplicative interaction parameter through a refined simulation based strategy by simulating from independent beta variates, the latter obtained by considering simple transformations on the main variates. The work of [16] invited attention of genetic epidemiologists; it however failed to provide any concrete suggestion on the choice of Dirichlet hyperparameters used in the study. The authors simply advocated for symmetric arbitrary choices based on prior strength but without any focussed and convincing justification for the same. Dirichlet prior is undoubtedly a flexible and a rich family but always problematic with regard to the choice of its hyperparameters. As such, the researchers finally recommend either symmetric choices or non-informative choice as a special case, leaving behind an important issue of why then to use Dirichlet prior. It appears as if it is simply because of its conjugacy with multinomial likelihood that, in turn, provides the ease of drawing inferences. We do not find any systematic work on the choice of hyperparameters for such a prior, an issue that always troubles the applied practitioners.

II. Manuscript Organization

The plan of the paper is as follows. The next section provides the necessary modelling formulation based on multinomial-Dirichlet assumption for a simple case-control setup classified according to disease status, gene status and environmental exposure components. We have extended the binary setup by allowing more than two categories on one of the classifying components. The various parameters of interest in our study of gene-environment interaction are also detailed. Section IV comments on the choice of hyperparameters and focuses, in particular, on symmetric choices. Since the Dirichlet prior offers a variety of shapes for the variation in its parameters, it is important to consider technique(s) for the choice of prior hyperparameters so that one may suggest an appropriately chosen prior for the study. EPPL criterion is also detailed in a separate subsection for completeness. This criterion is used to provide an additional weight on our study of choosing the hyperparameters appropriately.

Section V provides a description of data set based on case-control study of ovarian cancer patients in Israel. This data set reports observations on genetic and environmental components besides a few usual reporting about the patients. Moreover, the data set is a redesigned version of a real data set initially reported by [14] and reanalyzed by [4] and [15] among others. We, however, could not get the actual data due to confidentiality reason and, therefore, used the version that was reported by [4] on their web page. This reported version is based on a few actual and a few simulated observations (see Section V for details). We have also used an extracted data set of reduced size so that the procedure can be equally well illustrated for a comparatively smaller sample size. Section VI provides numerical illustration for the data sets given in Section V. The section aims to draw inferences on relevant model parameters. Results are provided for both binary and extended binary classifications. A brief conclusion and references are given at the end.

TABLE I
Classification of a case-control structure with respect to disease status, gene status and environmental exposure

	$G=0$				$G=1$				Total
	$E=0$	$E=1$	...	$E=l$	$E=0$	$E=1$	...	$E=l$	
$D=0$	r_{000}	r_{001}	...	r_{00l}	r_{010}	r_{011}	...	r_{01l}	n_0
$D=1$	r_{100}	r_{101}	...	r_{10l}	r_{110}	r_{111}	...	r_{11l}	n_1

III. Model Formulation

Let us consider a simple $2 \times 2 \times (l+1)$ setup in which D stands for the disease status with $D=0$ denoting a control and $D=1$ denoting a case. In a similar way, suppose the genetic susceptibility $G=0$ represents a non-carrier, $G=1$, a carrier of a particular genotype, $E=0$, an unexposed individual and $E=i$ represents an exposed individual with the level of environmental exposure i, $i=1,...,l$. The complete structure consisting of n (= n_0 + n_1) individuals is reported in Table I in the form of counts. These counts are represented as r_{DGE}, where D and G are binary variables taking values either 0 or 1 and E is a polytomous variable taking values $0,1,...,l$ for each combination of D and G.

This structure can be very well represented by a multinomial distribution with $\mathbf{r_0} \sim$ multinomial $(n_0, \mathbf{p_0})$ where $\mathbf{r_0} = (r_{000},...,r_{00l},r_{010},...,r_{01l})$ and $\mathbf{p_0} = (p_{000},...,p_{00l},p_{010},...,p_{01l})$. Similarly, $\mathbf{r_1} \sim$ multinomial $(n_1, \mathbf{p_1})$ where $\mathbf{r_1} = (r_{100},...,r_{10l},r_{110},...,r_{11l})$ and $\mathbf{p_1} = (p_{100},...,\ p_{10l},p_{110},...,p_{11l})$. The components of $\mathbf{p_0}$ and $\mathbf{p_1}$ are the corresponding cell probabilities, that is,

$$p_{0GE} = r_{0GE}/n_0, \text{ for } G=0,1; E=0,1,...,l,$$
$$p_{1GE} = r_{1GE}/n_1, \text{ for } G=0,1; E=0,1,...,l,$$

where $\Sigma r_{0GE} = n_0$ and $\Sigma r_{1GE} = n_1$. The likelihood function for cases and controls based on the above configuration can be written as

$$L_{CO}(\mathbf{p_0}) \propto {p_{000}}^{r_{000}}...{p_{00l}}^{r_{00l}}{p_{010}}^{r_{010}}...{p_{01l}}^{r_{01l}}, \quad (1)$$

and

$$L_{C}(\mathbf{p_1}) \propto {p_{100}}^{r_{100}}...{p_{10l}}^{r_{10l}}{p_{110}}^{r_{110}}...{p_{11l}}^{r_{11l}}, \quad (2)$$

respectively. Obviously, $\Sigma_{i=0}^{l}(p_{00i} + p_{01i}) = 1$ and $\Sigma_{i=0}^{l}(p_{10i} + p_{11i}) = 1$.

Let us now consider the Dirichlet priors for the corresponding cell probabilities $\mathbf{p_0} = (p_{000}, ..., p_{00l}, p_{010}, ..., p_{01l})$ and $\mathbf{p_1} = (p_{100}, ..., p_{10l}, p_{110}, ..., p_{11l})$ as

$$g_{CO}(\mathbf{p_0}|\mathbf{a}) \propto {p_{000}}^{a_0-1}...{p_{00l}}^{a_l-1}{p_{010}}^{a_{l+1}-1}...{p_{01l}}^{a_{2l+1}-1}, \tag{3}$$

and

$$g_{C}(\mathbf{p_1}|\mathbf{b}) \propto {p_{100}}^{b_0-1}...{p_{10l}}^{b_l-1}{p_{110}}^{b_{l+1}-1}...{p_{11l}}^{b_{2l+1}-1}, \tag{4}$$

where $\mathbf{a} = (a_0, a_1, ..., a_{2l+1})$ and $\mathbf{b} = (b_0, b_1, ..., b_{2l+1})$ are the hyperparameters. Thus to complete the Bayesian modelling formulation for the present case-control structure, we need to combine the priors in (3) and (4) with the likelihoods in (1) and (2), respectively, via Bayes' theorem. The corresponding posteriors up to proportionality can be written as

$$P_{CO}(\mathbf{p}_0 \mid \mathbf{a}, \mathbf{r}_0) \propto p_{000}^{r_{000}+a_0-1} \cdots p_{00l}^{r_{00l}+a_l-1} p_{010}^{r_{010}+a_{l+1}-1} \cdots p_{01l}^{r_{01l}+a_{2l+1}-1}. \tag{5}$$

and

$$P_{C}(\mathbf{p}_1 \mid \mathbf{b}, \mathbf{r}_1) \propto p_{100}^{r_{100}+b_0-1} \cdots p_{10l}^{r_{10l}+b_l-1} p_{110}^{r_{110}+b_{l+1}-1} \cdots p_{11l}^{r_{11l}+b_{2l+1}-1}. \tag{6}$$

Obviously, the posteriors (5) and (6), when normalized, are again Dirichlet distributions with updated parameters $(\mathbf{r_0} + \mathbf{a})$ and $(\mathbf{r_1} + \mathbf{b})$, respectively, and because of being available in closed forms, these can be easily generated by a number of techniques (see, for example, Devroye (1986)).

Let $OR_{10_i} = p_{000}p_{10i}/p_{00i}p_{100}$ denotes the odds ratio associated with $E = i$ for non-susceptible subjects $(G = 0)$ and $OR_{01} = p_{000}p_{110}/p_{010}p_{100}$ denotes the odds ratio associated with G for unexposed individuals $(E = 0)$. Similarly, the odds ratio associated with $G = 1$ and $E = i$ compared to $G = 0$ and $E = 0$ is denoted by $OR_{11_i} = p_{000}p_{11i}/p_{01i}p_{100}$, $i = 1, ..., l$. The measure of $G - E$ association in the control population at i^{th} level of E is, therefore, given by,

$$\theta_{GE_i} = log(p_{000}p_{01i})/(p_{00i}p_{010}), i = 1, ..., l. \tag{7}$$

An easy procedure for getting the sample based estimates of θ_{GE_i} and other log odds ratios is straightforward. We simply need to simulate the values of $\mathbf{p_0}$ and $\mathbf{p_1}$ from the posterior distributions (5) and (6), respectively, and after substituting the generated $\mathbf{p_0}$ and $\mathbf{p_1}$ in the concerned relationships, we may obtain the corresponding samples from the posterior distributions of the same. These samples may be used to estimate the entire posterior distribution of the corresponding variate, say, using kernel density estimate or histogram, etc. One may also draw other desired features of interest in a routine manner. Suppose, for example, one is interested in the point estimate of θ_{GE_i} and finds the estimated value close to zero. This means that G and $E = i$ are not dependent on each other in the control population. Contrary to that if the estimated θ_{GE_i} is non-zero, one may go a step ahead and calculate the multiplicative interaction parameter between G and $E = i$ [15] given by,

$$\psi_i = (p_{00i}p_{010}p_{100}p_{11i})/(p_{000}p_{01i}p_{10i}p_{110}), i = 1, ..., l. \tag{8}$$

(8) provides a measure of association between the gene and environmental component $E = i, i = 1, ..., l$, and this may be the parameter where the interest often centers. Moreover, the measures of association and interaction as given in (7)-(8) are calculated fixing the base environmental exposure $E = 0$ although such measures can be obtained among any pairs of environmental exposure $E = i$ and $E = j$, $i \neq j \neq 0$, provided the interest centres among different pairs and the cell frequencies are large enough to support such evaluations.

IV. Choice of Hyperparameters

As pointed out earlier, the Dirichlet prior offers a conjugate family for multinomial likelihood and, as such, the family is quite flexible and rich and it results in computationally easy inferences. The problem, however, arises with the choice of hyperparameters and it increases its severity with the increasing dimensionality of the entertained Dirichlet model. To choose the hyperparameters of Dirichlet prior in our problem, we shall follow the following strategy here:

A. Symmetric Dirichlet Prior

When we consider equal values of different components of vector valued parameter of a Dirichlet prior, the resulting distribution may be called symmetric Dirichlet distribution. Symmetric Dirichlet distributions are often used as a prior when there is no particular preference of one component over other resulting into what we may call as weak or vague prior. Since all the components of the vector parameter have the same value, the distribution can alternatively be written by means of a single scalar parameter called the concentration parameter. One can, of course, entertain symmetric Dirichlet distribution as a prior but again the question arises whether one should go with small or large choices of hyperparameters as such choices highly affect the concentration of the model towards its mean value. Say, for instance, if all the components of hyperparameters **a** and **b** are taken to be unity, the priors (3) and (4) reduce to uniform priors in the range (0, 1). Such priors may be considered non-informative as suggested by [9]. Similarly, if all the components of **a** and **b** are taken to be zero, the resulting priors may be treated as vague although such choices are strictly outside the parameter space (see [12]). Values of the concentration parameter above unity prefer variates that are dense and evenly-distributed. Values of the concentration parameter below unity prefer sparse distributions, that is most of the probabilities returned will be close to zero and the vast majority of the mass will be concentrated in a few of the probabilities. A systematic discussion on Dirichlet prior and the choices of its hyperparameters, especially when the entertained model is symmetric, can be had from [6]. For the situations when Dirichlet model is desired to be non-symmetric, the choice of hyperparameters can be even more problematic.

Here we propose to consider a symmetric Dirichlet prior and arbitrarily consider a few equal sets of values for the hyperparameters. We shall finally use EPPL criterion (see also [3]) to compare the various modelling combinations in order to see if a particular set of hyperparameters may be considered better than all others. It is important to mention that the EPPL criterion is often recommended for comparing two or more models based on their goodness of fit and complexity. This criterion may not be sensitive enough for recommending a particular prior corresponding to a particular set of hyperparameter combinations but the issue may not be detrimental as we have a very large sample size in the present case. We shall provide below a brief review of the EPPL criterion. This description has been given for completeness only.

B. Expected Posterior Predictive Loss Criterion

Predictive loss criterion for model selection was proposed by [5] (see also [10]). The criterion utilizes both the observed and the predictive data sets and advocates for a model based on minimum posterior predictive loss. Use of predictive distributions for model selection appears convincing in the sense that it is natural to consider the model performance by comparing what it predicts with what it observes. The authors consider the standard utility ideas as advocated by [18] but replacing experiments with the models and minimizing loss over the models. We feel as if the criterion proposed by [5] is perhaps the most convincing criterion since it involves predictive samples which are the results of both likelihood and prior combination although it appears that the measure may not be sensitive enough for the variations in prior components only.

The posterior predictive loss criterion of [5] was later modified by [3] so as to include count data given in the form of multinomial cell frequencies or contingency tables. The authors referred the criterion as the expected predictive deviance though it is same as the expected value of the loss function under the posterior predictive distribution. We shall call it as the EPPL criterion.

The criterion, in fact, comprises a component known as the likelihood ratio statistic (LRS), a quantity signifying goodness of fit, and a component that may be interpreted as the penalty (PEN) for model complexity. Its general form can be written as (see, for example, [3]),

$$EPPL = E[L(\mathbf{x}_{rep}, \mathbf{x}_{obs})] = LRS + PEN, \qquad (9)$$

where $\mathbf{x}_{obs}$ ($\mathbf{x}_{rep}$) denotes the observed (predictive) data with individual cell data counts $x_{i,obs}$ ($x_{i,rep}$), $i = 1, ..., k$, and k is the number of classes. To avoid difficulty we can replace the zero counts by adding $1/2$ in that particular cell. The loss function $L(\mathbf{x}_{rep}, \mathbf{x}_{obs})$, called the deviance loss function, is defined as

$$L(\mathbf{x}_{rep}, \mathbf{x}_{obs}) = 2\Sigma_i x_{i,obs} \log\left(\frac{x_{i,obs}}{x_{i,rep}}\right). \qquad (10)$$

Taking expectation of this loss function, $EPPL$ and LRS are obtained respectively as,

$$EPPL = 2\Sigma_i x_{i,obs}[\log x_{i,obs} - E(\log x_{i,rep})], \qquad (11)$$

and

$$LRS = 2\Sigma_i x_{i,obs}[\log(x_{i,obs}/n) - \log(p_i^*)], \qquad (12)$$

where the expectation is taken with respect to the posterior predictive distribution

$$\pi(\mathbf{x}_{rep}|\mathbf{x}_{obs}) = \int f(\mathbf{x}_{rep}|\mathbf{p}) h(\mathbf{p}|\mathbf{x}_{obs}) d\mathbf{p}. \qquad (13)$$

Here n is the total cell observations of k classes, $\mathbf{p}$ stands for the parameter vector with components p_i denoting the probability of i^{th} class and $p_i^* = E(x_{i,rep}/n)$, $i = 1, ..., k$ (see also [3]). Both $EPPL$ and LRS can be easily obtained from (11) and (12), respectively, using simulated predictive samples and then PEN can be obtained from (9) by subtraction.

Clearly, $EPPL$ consists of two parts. The first one is a goodness of fit term, which gives the loss incurred due to incompatibility of the model to perform at observed data points. The other term is the penalty term associated with the model complexity and it increases with the variability of $\mathbf{x}_{rep}$. Thus, the EPPL criterion recommends a model (that is, a prior-likelihood combination) that has the smallest loss among all the competing ones.

V. Data Description

We consider a partially real and partially simulated data set on Jewish women in Israel which was taken from the web page of [4]. The necessary details can be had from http://dceg.cancer.gov/people/ Chatterjee Nilanjan.html under the software link. This data set consists of real data values on disease status, D, and non-genetic co-factors, Y. For reasons of privacy, however, the real genetic data are not available publicly. Instead, the data consist of simulated genetic data, G, generated using the conditional distribution of $[G|D, Y]$ as specified by the parameter estimates obtained from the real data (see [4] for details).

In spite of the fact that multiparity and use of OC reduce the risk of ovarian cancer in women, we propose to study the effect of these factors on women with BRCA 1 and/or BRCA 2 mutation as well. The data set contains 747 controls and 832 cases of women who underwent mutation analysis. The data set, referred to as the case-control data, is given in Table II for a ready reference.

For OC use, $E = 0$ corresponds to those subjects who never used OC. $E = i$ corresponds to those who used OC up to 3 years, from 3 years to 6 years and for more than 6 years depending on whether $i = 1, 2$, and 3, respectively. Similarly, for parity, $E = 0$ corresponds to women who have no children, $E = i$ corresponds to 1-3 children, 3-6 children, and more than 6 children accordingly as $i = 1, 2$, and 3, respectively.

Above data can be easily converted in the form of a $2\times2\times2$ design by combining cells with non-zero E in to a single cell, say, $E = 1$. Therefore, in this case, when OC use is considered as the environmental exposure, there are 586 unexposed individuals among controls. That is, among 747

TABLE II
CLASSIFICATION OF CASE-CONTROL DATA WITH RESPECT TO DISEASE STATUS, GENETIC SUSCEPTIBILITY AND ENVIRONMENTAL EXPOSURE

	OC Use								
	$G=0$				$G=1$				Total
	$E=0$	$E=1$	$E=2$	$E=3$	$E=0$	$E=1$	$E=2$	$E=3$	
$D=0$	577	86	32	40	9	1	1	1	747
$D=1$	494	67	15	16	184	34	7	15	832
	Parity								
$D=0$	42	506	155	32	1	8	2	1	747
$D=1$	68	373	116	35	20	188	30	2	832

controls, 586 are those who are not using OC whereas 161 individuals are using the same. While among cases, 678 are unexposed to environment and 154 are exposed. Similarly, when parity is considered as the environmental exposure, only 43 individuals are unexposed among controls while 704 are exposed. Among cases the number of unexposed individuals is 88 while that of exposed is 744.

We also considered a data of moderate sample size with $n = 50$ extracted from $2 \times 2 \times 2$ setup discussed above in such a way that each cell frequency remains at least equal to unity. The data set is shown in Table III and it is taken exclusively for the purpose of comparison. It is to be noted that such data sets with small to very small sample sizes may not result in the desired inferences on association and interaction parameters as the corresponding reported cell frequencies may not be the true representatives of the prevalence of gene and environmental components in the actual population.

TABLE III
CLASSIFICATION OF EXTRACTED DATA ($2 \times 2 \times 2$) OF SIZE 50 WITH RESPECT TO DISEASE STATUS, GENETIC SUSCEPTIBILITY AND ENVIRONMENTAL EXPOSURE

	OC Use				
	$G=0$		$G=1$		Total
	$E=0$	$E=1$	$E=0$	$E=1$	
$D=0$	19	3	1	1	24
$D=1$	16	2	5	3	26
	Parity				
$D=0$	2	20	1	1	24
$D=1$	2	15	2	7	26

VI. NUMERICAL ILLUSTRATION

For the purpose of numerical illustration, we consider the same data set reported earlier in Section V and analyze on the basis of modelling formulation given in Section III. The primary objective is to estimate the odds ratio, measure of association between G and E and, in case this association is non-zero, the multiplicative interaction parameter between the latter two. Since the considered prior involves a number of hyperparameters, we shall consider the choice suggested in the subsection IV-A for choosing the same. It is to be noted, however, that the $G-E$ association in the control population at i^{th} level of E (see (7)) and the multiplicative interaction parameter between G and $E = i$, $i = 1, 2, 3$ (see (8)) based on the observations in Table II are not logically appealing as most of the cell frequencies are too small to draw any valid conclusion on the actual $G-E$ association or the multiplicative interaction parameter. We, therefore, propose to consider such measures for $2 \times 2 \times 2$ setup also (see also Section V) along with the $2 \times 2 \times (l+1)$ setup. In the former case, we shall drop the subscript i from OR_{10_i}, θ_{GE_i}, and ψ_i keeping their interpretations same.

A. Results Based on Symmetric Dirichlet Prior

At first, we obtain the point estimates of different cell probabilities $\mathbf{p_0}$ and $\mathbf{p_1}$ in the form of posterior means. Although closed form expressions for these estimates can be obtained directly from (5) and (6), we used samples from (5) and (6) instead and considered the sample based estimates. Moreover, these samples can be used to estimate other features of $\mathbf{p_0}$ and $\mathbf{p_1}$ as well but we are confined to posterior means only for the purpose of illustration. For priors, we used the symmetric Dirichlet distributions by assigning a single scalar value to all the hyperparameter components. Since the values above unity prefer variates that are dense, evenly-distributed distributions, that is, all probabilities returned are similar to each other, we advocate using the values greater than or equal to unity for all the hyperparameter components. The results based on different symmetric choices of prior hyperparameters are shown in Table IV corresponding to the observations given in Table II. These estimates, based on posterior samples of size 5000 from each, are obtained separately for both OC use and parity (see Table IV). It is to be noted that the constants '**a**' and '**b**' are not the scalars but the vectors with $\mathbf{a} = (a_0, a_1, a_2, a_3, a_4, a_5, a_6, a_7)$ and $\mathbf{b} = (b_0, b_1, b_2, b_3, b_4, b_5, b_6, b_7)$ and $a = v_1$, $b = v_2$ imply that all the components of **a** are v_1 and those of **b** are v_2.

It can be seen from the results that the estimated cell probabilities are not affected much for different equal choices of hyperparameter components in spite of the fact that a number of observed cell frequencies are either unity or close to unity. The reason might be attributed to the fact that we have considerably large overall sample size and, as a result, the likelihood becomes dominating over the prior distribution. We have also obtained the standard deviations of the posterior samples shown in parentheses. These estimates are based on a sample of size 5000 from the corresponding marginal posterior. As expected these estimated standard deviations are also least affected by the changes in the prior hyperparameter components. Since both the estimates and the estimated standard deviations of cell probabilities are least affected for the symmetric variation in the hyperparameters, one can consider the hyperparameter combinations which result in exactly uniform Dirichlet kernel (all the components unity). These choices will at least make a sense that no *a priori* preference is given to any of the cell probabilities in the absence of any prior information.

Although we have advocated for the use of symmetric Dirichlet prior with all the hyperparameter components set to unity, it is vital to consider some sort of comparison before affirming this choice. As discussed, we consider evaluation of

TABLE IV
ESTIMATED SAMPLE BASED POSTERIOR MEANS AND THE CORRESPONDING STANDARD DEVIATIONS OF DIFFERENT CELL PROBABILITIES BASED ON SYMMETRIC CHOICES

Cell	$a=1.0, b=1.0$		$a=2.0, b=2.0$		$a=5.0, b=5.0$		$a=10.0, b=10.0$	
probs.	OC use	Parity	OC use	Parity	OC use	Parity	OC use	Parity
p_{000}	0.7648	0.0567	0.7582	0.0574	0.7389	0.0594	0.7092	0.0626
	(0.0147)	(0.0082)	(0.0147)	(0.0082)	(0.0148)	(0.0082)	(0.0153)	(0.0083)
p_{001}	0.1156	0.6718	0.1157	0.6661	0.1161	0.6496	0.1165	0.6243
	(0.0108)	(0.0164)	(0.0108)	(0.0164)	(0.0106)	(0.0164)	(0.0104)	(0.0163)
p_{002}	0.0442	0.2069	0.0451	0.2061	0.0475	0.2036	0.0512	0.1999
	(0.0074)	(0.0143)	(0.0074)	(0.0143)	(0.0075)	(0.0139)	(0.0075)	(0.0134)
p_{003}	0.0544	0.0438	0.0552	0.0446	0.0573	0.0471	0.0606	0.0508
	(0.0086)	(0.0078)	(0.0086)	(0.0079)	(0.0086)	(0.0079)	(0.0087)	(0.0080)
p_{010}	0.0131	0.0025	0.0142	0.0038	0.0176	0.0075	0.0228	0.0131
	(0.0040)	(0.0017)	(0.0042)	(0.0021)	(0.0046)	(0.0029)	(0.0051)	(0.0038)
p_{011}	0.0029	0.0121	0.0041	0.0132	0.0077	0.0166	0.0134	0.0218
	(0.0021)	(0.0043)	(0.0025)	(0.0044)	(0.0033)	(0.0049)	(0.0043)	(0.0055)
p_{012}	0.0024	0.0037	0.0037	0.0050	0.0075	0.0088	0.0132	0.0144
	(0.0017)	(0.0021)	(0.0021)	(0.0024)	(0.0028)	(0.0030)	(0.0036)	(0.0038)
p_{013}	0.0026	0.0025	0.0038	0.0038	0.0074	0.0074	0.0131	0.0131
	(0.0018)	(0.0018)	(0.0022)	(0.0022)	(0.0031)	(0.0031)	(0.0041)	(0.0041)
p_{100}	0.5904	0.0826	0.5860	0.0830	0.5733	0.0842	0.5537	0.0860
	(0.0161)	(0.0090)	(0.0161)	(0.0090)	(0.0159)	(0.0089)	(0.0156)	(0.0088)
p_{101}	0.0820	0.4469	0.0824	0.4439	0.0836	0.4351	0.0854	0.4216
	(0.0091)	(0.0171)	(0.0091)	(0.0171)	(0.0091)	(0.0167)	(0.0089)	(0.0163)
p_{102}	0.0192	0.1392	0.0202	0.1391	0.0231	0.1387	0.0276	0.1381
	(0.0049)	(0.0123)	(0.0049)	(0.0122)	(0.0052)	(0.0120)	(0.0055)	(0.0117)
p_{103}	0.0198	0.0421	0.0208	0.0429	0.0236	0.0451	0.0280	0.0486
	(0.0048)	(0.0069)	(0.0049)	(0.0069)	(0.0052)	(0.0070)	(0.0055)	(0.0071))
p_{110}	0.2186	0.0246	0.2178	0.0254	0.2152	0.0282	0.2113	0.0324
	(0.0142)	(0.0052)	(0.0142)	(0.0053)	(0.0138)	(0.0055)	(0.0134)	(0.0057)
p_{111}	0.0415	0.2244	0.0423	0.2235	0.0446	0.2208	0.0481	0.2166
	(0.0066)	(0.0137)	(0.0066)	(0.0136)	(0.0067)	(0.0134)	(0.0068)	(0.0130)
p_{112}	0.0096	0.0368	0.0106	0.0377	0.0138	0.0401	0.0187	0.0438
	(0.0034)	(0.0066)	(0.0036)	(0.0066)	(0.0040)	(0.0067)	(0.0046)	(0.0069)
p_{113}	0.0189	0.0034	0.0199	0.0045	0.0228	0.0078	0.0272	0.0129
	(0.0046)	(0.0019)	(0.0047)	(0.0022)	(0.0050)	(0.0029)	(0.0053)	(0.0037)

$EPPL$s (see subsection IV-B) for various entertained model-prior combinations. It is to be noted that our model is fixed here and the only change in prior is because of its hyperparameters. A remark: we understand that the EPPL criterion is actually meant for comparing the models and it might not be a sensitive criterion for comparing the variations in hyperparameters, we still use it because the model comparison in a Bayesian framework checks for the entire package that includes both the likelihood and the prior combinations. We feel as if the criterion must provide extra weight to our already arrived conclusion.

The corresponding results for the $EPPL$s are shown in Table V. It can be seen from the results that the values of $EPPL$ are minimum for the hyperparameter combination $a = 1.0$ and $b = 1.0$, hence this choice may be considered as the most appropriate choice for both OC use and parity, a conclusion that was suggested earlier too. Contrary to our expectation, we see slight variation in the values of PEN as well although the model complexity remains same for changing hyperparameters. It seems as if the change in the values of hyperparameters changes the posterior samples as well as the predictive samples, the latter causes the change in predictive variability as well.

We finally consider the estimated posterior means and standard deviations of log odds ratio and other interactive

TABLE V
$EPPL$**S CORRESPONDING TO VARIOUS CHOICES OF HYPERPARAMETERS**

Environmental Exposure	Hyperparameter Values	Case control	Loss due to fitting	Loss due to penalty	$EPPL$
OC use	$a=1.0, b=1.0$	Control	1.28	4.38	5.66
		Case	0.38	4.70	5.08
	$a=2.0, b=2.0$	Control	2.73	3.81	6.54
		Case	0.40	4.78	5.18
	$a=5.0, b=5.0$	Control	8.43	2.88	11.31
		Case	0.47	4.80	5.27
	$a=10.0, b=10.0$	Control	20.06	2.86	22.92
		Case	0.63	4.95	5.58
Parity	$a=1.0, b=1.0$	Control	1.36	3.34	4.70
		Case	0.40	4.42	4.82
	$a=2.0, b=2.0$	Control	3.14	3.12	6.26
		Case	0.52	4.37	4.89
	$a=5.0, b=5.0$	Control	9.35	2.46	11.81
		Case	1.05	4.23	5.28
	$a=10.0, b=10.0$	Control	22.84	2.37	25.21
		Case	2.1	3.92	6.02

parameters (see Section III) based on samples of size 5000. These estimates are based on both $2\times2\times2$ and $2\times2\times4$ setup for the reasons mentioned earlier. Moreover, the samples from these different characteristics are obtained from the samples of $\mathbf{p_0}$ and $\mathbf{p_1}$ merely by substitution. The estimated posterior means and the corresponding standard deviations (in parentheses) for hyperparameter combination $a = 1.0$ and $b = 1.0$ are shown in Table VI and VII. It is important to mention that the posterior samples of log odds ratios and other interactive parameters can also be used to estimate several other posterior features including even the density estimates but we are not giving those features presuming as if these are routine tasks.

TABLE VI
POSTERIOR ESTIMATES OF LOG ODDS RATIO AND OTHER INTERACTIVE PARAMETERS FOR $2\times2\times2$ **SETUP**

Parameters	OC use	Parity
θ_{GE}	0.2713	-0.7666
	(0.6474)	(0.8841)
$\log(OR_{10})$	-0.3199	-0.7538
	(0.1494)	(0.2099)
$\log(OR_{01})$	3.1147	2.1346
	(0.3343)	(0.8808)
$\log(\psi)$	0.1621	1.1169
	(0.6716)	(0.9340)

It can be seen that the estimated value of θ_{GE} is different from zero for both OC use and parity conveying that genetic susceptibility and environmental exposures could not be considered independent as per our presumption given at the beginning. $\log(OR_{10})$ is coming out to be negative for every case representing that for non-susceptible subjects both OC use and parity reduce the risk of ovarian cancer. On the other hand, positive values of $\log(OR_{01})$ suggest that the subjects unexposed to environmental factors but having BRCA1/2 mutation are having a greater risk of ovarian cancer. The estimated ψ is showing strong positive association among genetic susceptibility and environmental exposure components

TABLE VII
POSTERIOR ESTIMATES OF LOG ODDS RATIO AND OTHER INTERACTIVE PARAMETERS FOR $2 \times 2 \times 4$ SETUP

Posterior estimates	OC use	Parity	Posterior estimates	OC use	Parity
θ_{GE_1}	0.1536	-0.7516	$\log(OR_{10_3})$	-0.7679	-0.4161
	(0.8991)	(0.8416)		(0.2937)	(0.3139)
θ_{GE_2}	0.9495	-0.8356	$\log(OR_{01})$	3.1193	2.1026
	(0.8540)	(0.9482)		(0.3384)	(0.7960)
θ_{GE_3}	0.7893	0.2295	$\log(\psi_1)$	0.1537	1.2936
	(0.9004)	(1.1447)		(0.9162)	(0.8748)
$\log(OR_{10_1})$	-0.0862	-0.7893	$\log(\psi_2)$	-0.6843	0.7262
	(0.1652)	(0.1988)		(0.9767)	(1.0153)
$\log(OR_{10_2})$	-0.5929	-0.7792	$\log(\psi_3)$	0.1553	-1.6712
	(0.3156)	(0.2214)		(0.9740)	(1.3652)

whether latter is concerned with OC use or parity. Obviously, this strong positive association can be viewed as an important risk for the development of ovarian cancer. Moreover, this association is quite high when parity is taken as the environmental exposure component (see Table VI). The estimated standard deviations are quite small in each case (see also Table IV) giving a clear-cut impression that the estimates, in general, are highly stabilized.

Observing Table VII, we can conclude that $\theta_{GE_i}, i = 1, ..., 3$, is clearly showing association for both OC use and parity discarding the assumption of independence between genetic susceptibility and environmental exposures. $log(OR_{10_i}), i = 1, ..., 3$, is as usual coming out to be negative suggesting that for non-susceptible subjects both OC use and parity reduce the risk of ovarian cancer and this risk mostly decreases for higher levels of OC use and/or parity. The values of $\log(OR_{01})$ again help us to reach the conclusion that the patients with BRCA1/2 mutation are having high risk of developing ovarian cancer in control population. Moreover, the values of interaction parameters are conveying a very important message too. The interaction between BRCA 1/2 and OC use definitely increases the risk of ovarian cancer, however, this risk is reduced as compared to the situation when BRCA1/2 was acting alone. The negative value of $\log(\psi_2)$ is somewhat odd perhaps due to the small cell frequencies. Coming to the interaction between BRCA1/2 and parity, we can see that the risk decreases as the parity increases.

B. Results for an Extracted Data of Reduced Size

The results were also obtained for an extracted data set of size 50 given in Table III. This was done to see the effect of reduced sample size on our analysis. It is important to mention that if we have very small sample sizes, we may come across a situation where most of the cell frequencies are zero especially for large values of l. We, therefore, did not consider a value of n smaller than 50. The estimated cell probabilities along with their standard deviations (in parentheses) are given in Table VIII. Contrary to our expectation, we noted that the estimated cell probabilities were showing more or less similar trend that we observed earlier for $2 \times 2 \times 2$ setup.

We finally calculated the posterior estimates of log odds ratio and other interactive parameters which are shown in Table

TABLE VIII
ESTIMATED SAMPLE BASED POSTERIOR MEANS OF DIFFERENT CELL PROBABILITIES FOR EXTRACTED DATA OF SIZE 50

Cell	$a = 1.0, b = 1.0$		$a = 2.0, b = 2.0$		$a = 5.0, b = 5.0$		$a = 10.0, b = 10.0$	
probs.	OC use	Parity	OC use	Parity	OC use	Parity	OC use	Parity
p_{000}	0.7180	0.1113	0.6611	0.1294	0.5501	0.1635	0.4568	0.1913
	(0.0855)	(0.0604)	(0.0852)	(0.0613)	(0.0772)	(0.0587)	(0.0644)	(0.0522)
p_{001}	0.1407	0.7471	0.1545	0.6859	0.1809	0.5670	0.2026	0.4678
	(0.0639)	(0.0827)	(0.0628)	(0.0829)	(0.0583)	(0.0757)	(0.0511)	(0.0634)
p_{010}	0.0699	0.0701	0.0921	0.0922	0.1357	0.1359	0.1722	0.1723
	(0.0465)	(0.0464)	(0.0501)	(0.0499)	(0.0508)	(0.0505)	(0.0464)	(0.0462)
p_{011}	0.0714	0.0715	0.0923	0.0925	0.1333	0.1336	0.1684	0.1686
	(0.0484)	(0.0479)	(0.0514)	(0.0509)	(0.0514)	(0.0514)	(0.0467)	(0.0469)
p_{100}	0.5669	0.0985	0.5296	0.1166	0.4572	0.1521	0.3950	0.1824
	(0.0847)	(0.0530)	(0.0802)	(0.0527)	(0.0689)	(0.0495)	(0.0561)	(0.0442)
p_{101}	0.1011	0.5332	0.1190	0.4999	0.1532	0.4347	0.1819	0.3784
	(0.0532)	(0.0902)	(0.0543)	(0.0852)	(0.0527)	(0.0730)	(0.0478)	(0.0599)
p_{110}	0.1977	0.1005	0.2034	0.1179	0.2150	0.1514	0.2254	0.1807
	(0.0707)	(0.0547)	(0.0672)	(0.0554)	(0.0595)	(0.0530)	(0.0510)	(0.0476)
p_{111}	0.1343	0.2679	0.1480	0.2656	0.1746	0.2618	0.1977	0.2585
	(0.0632)	(0.0822)	(0.0620)	(0.0775)	(0.0571)	(0.0664)	(0.0503)	(0.0550)

TABLE IX
POSTERIOR ESTIMATES OF LOG ODDS RATIO AND OTHER INTERACTIVE PARAMETERS FOR EXTRACTED DATA OF SIZE 50

Parameters	OC use	Parity
θ_{GE}	1.7552	-2.0417
	(1.2625)	(1.3085)
$\log(OR_{10})$	-0.1236	-0.2335
	(0.8844)	(0.9523)
$\log(OR_{01})$	1.4676	0.5601
	(1.0050)	(1.3981)
$\log(\psi)$	0.3356	1.3109
	(1.6073)	(1.6695)

IX. It can be seen that the results are certainly changed in terms of numerical values but the final messages are more or less same that we drew earlier based on $2 \times 2 \times 2$ setup for large sample size. That is, θ_{GE} is showing an association among genetic and environmental components for both OC use and parity indicating the need of going a step ahead for evaluating the multiplicative interaction parameter ψ. Similarly, for non-susceptible subjects both OC use and parity reduce the risk of ovarian cancer whereas the subjects unexposed to environmental factors but having BRCA1/2 mutation are having a higher risk of the disease. The estimated ψ is also showing a similar behaviour with values indicating that OC use and parity reduce the risk of ovarian cancer even in patients with BRCA1/2 mutation. This reduction is stronger for OC use than for parity.

VII. CONCLUSION

This work is a successful attempt to study the interaction between genetic susceptibility and environmental exposure components as important causes for the development of ovarian cancer in women based on the assumption of multinomial-Dirichlet modelling. Other parameters involved in the modelling process are also estimated and their estimated standard deviations are obtained. Throughout we have used sample

based approaches because of their apparent advantages in the sense that every desired inference can be routinely obtained. This is otherwise difficult especially when one is concerned with functions of the parameters like odds ratio, multiplicative interaction, etc.

The second important finding relates to assessing the values of hyperparameters of the Dirichlet prior. Dirichlet prior is most widely recommended as a conjugate prior for multinomial likelihood. It has plenty of advantages, provides ease of drawing inferences but there appears to be no systematic discussion in the literature on choosing its hyperparameters. It is successfully shown that symmetric choices may be considered as a good alternative to select the hyperparameters. The EPPL criterion considered in the paper further strengthens our strategy for choosing the prior hyperparameters.

REFERENCES

[1] Arora, P., Upadhyay, S. K. and Shukla, V. K. (2011). A Bayesian case-control study to study the effect of covariates in gall bladder carcinoma. *STM Journals, Research & Reviews: A Journal of Medicine*, vol. 1(1), pp. 1–24.

[2] Ashby, D., Hutton, J. L. and McGee, M. A. (1993). Simple Bayesian analyses for case-control studies in cancer epidemiology. *The Statistician*, vol. 42, pp. 385–397.

[3] Buck, C.E. and Sahu, S.K. (2000). Bayesian models for relative archaeological chronology building. *Journal of Applied Statistics*, vol. 49, pp. 423–440.

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

[5] Gelfand, A. E. and Ghosh, S. K. (1998). Model Choice: A minimum posterior predictive loss approach. *Biometrika*, vol. 85, pp. 1–11.

[6] Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. New York: Springer-Verlag.

[7] Gupta, A. and Upadhyay, S.K. (2014). A Hierarchical Bayes Analysis to Study the Impact of Genetic Susceptibility and Environmental Exposure. *Research & Reviews: Discrete Mathematical Structures*, vol. 1(3), pp. 1-11.

[8] Gupta, A. and Upadhyay, S.K. (2019). Subjective Elicitation of Dirichlet Hyperparameters using Past Data – A Study of Ovarian Cancer Patients. *; Austrian Jr. Of Statistics*, vol. 48, pp. 1-14.

[9] Laplace, P. (1812). *Theorie Analytique Des Probabilities*. Paris: Courcier.

[10] Laud, P. W. and Ibrahim, J. G. (1995). Predictive model selection. *Journal of Royal Statistical Society B*, vol. 57, pp. 247–262.

[11] Lee, W. C., Wang, L. Y. and Cheng, K. F. (2010). An Easy-to-implement Approach for Analyzing Case-control and Case-only Studies Assuming Gene-environment Independence and Hardy-Weinberg Equilibrium. *Statistics in Medicine*, vol. 29, pp. 2557–2567.

[12] Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint, part 2 - Inference*. Cambridge: University Press.

[13] Makkar, P. (2009). *Bayesian Solutions of Some Medical Data Problems*. Unpublished Ph. D. Thesis, Banaras Hindu University, India.

[14] Modan, B., Hartge, P., Hirsh-Yechezkel, G., Chetrit, A., Lubin, F., Beller, U., Ben-Baruch, G., Fishman, A., Menczer, J., Struewing, J. P., Tucker, M. A. and Wacholder, S. (2001). Parity, oral contraceptives and the risk of ovarian cancer among carriers and non-carriers of a BRCA1 or BRCA2 mutation. *New England Jr. of Medicine*, vol. 345, pp. 235–240.

[15] 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*, vol. 64, pp. 685–694.

[16] Mukherjee, B., Ahn, J., Gruber, S. B., Ghosh, M. and Chatterjee, N. (2010). Case-control studies of gene-environment interaction: Bayesian design and analysis. *Biometrics*, vol. 66(3), pp. 934–948.

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

[18] Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Adminitration, Harvard University: Boston.

[19] Rothman, K. J. (1986). *Modern Epidemiology*. Little Brown and Company: Boston.