Semiparametric Bayesian Analysis of Matched Case-Control Studies With Missing Exposure

Samiran SINHA, Bhramar MUKHERJEE, Malay GHOSH, Bani K. MALLICK, and Raymond J. CARROLL

This article considers Bayesian analysis of matched case-control problems when one of the covariates is partially missing. Within the likelihood context, the standard approach to this problem is to posit a fully parametric model among the controls for the partially missing covariate as a function of the covariates in the model and the variables making up the strata. Sometimes the strata effects are ignored at this stage. Our approach differs not only in that it is Bayesian, but, far more importantly, in the manner in which it treats the strata effects. We assume a Dirichlet process prior with a normal base measure for the stratum effects and estimate all of the parameters in a Bayesian framework. Three matched case-control examples and a simulation study are considered to illustrate our methods and the computing scheme.

KEY WORDS: Case-control studies; Conditional inference; Dirichlet process; Endometrial cancer; Equine epidemiology; Exponential family; Low birth weight study; Matching; Metropolis–Hastings; Missing data; Retrospective studies.

1. INTRODUCTION

This article concerns matched case-control studies when one of the covariates is partially missing. Case-control studies are the dominant form of analytical research in epidemiology, with matched case-control studies having the advantage of being matched on the basis of important stratification variables. (See Breslow and Day 1980; Breslow 1996 for background and references.) With no missing data, there is a complementary Bayesian literature for ordinary and matched case-control studies (see Zelen and Parker 1986; Müller and Roeder 1997; Diggle, Morris, and Wakefield 2000; Seaman and Richardson 2001; Ghosh and Chen 2002).

In a matched analysis, some variables may not be observed for all study subjects. For example, in the Los Angeles study on endometrial cancer in a retirement community (Breslow and Day 1980), a binary variable denoting obesity was missing in approximately 16% of respondents. There have been two main approaches to handle the missingness in matched case-control problems. Lipsitz, Parzen, and Ewell (1998) and Rathouz, Satten, and Carroll (2002) modeled the missingness process directly. More germane to the present article is the likelihood approach that involves modeling the missing covariate distribution among the controls (see Satten and Kupper 1993a,b, who initiated this approach). Wang, Wang, and Carroll (1997) and Paik and Sacco (2000) attacked the missing-data problem directly by positing that the exposure distribution among the controls belongs to a canonical exponential family. Although the Paik and Sacco model is fully parametric, their estimation method is a pseudolikelihood methodology rather than full likelihood, and in the absence of missing data it reduces to the usual conditional logistic regression (CLR). Satten and Carroll (2000) generalized the Satten–Kupper and Paik–Sacco methodology to allow a full parametric likelihood analysis, allowing essentially any exposure distribution among the controls and providing for full likelihood analysis; their approach does not reduce to the usual CLR with no missing data.

The present article develops a Bayesian approach to case-control studies with a disease indicator D, a completely observed covariate vector $\mathbf{Z}$ and an exposure or risk factor X with possibly missing values. We also include variables $\mathbf{S} = (S_o, S_u)$, which define the strata used for matching. These are generally a vector of covariates, so that S_o denotes covariates that are explicit factors making up the strata and S_u denotes factors that are not explicitly used to form strata but nonetheless are associated with the strata. The variable X that is partially missing can be discrete or continuous.

An example will help clarify the strata issue. This example, specifically considered numerically in this article, is due to Kim, Cohen, and Carroll (2002). They described a 1:1 matched case-control study that investigated the association of various management practices with equine colic. Participating veterinarians were asked to provide data monthly for one horse that was treated for colic and one horse that received emergency treatment for any condition other than colic between March 1, 1997 and February 28, 1998. A case of colic was defined as the first horse treated during a given month for signs of intra-abdominal pain. A control horse was defined as the next horse that received emergency treatment for any condition other than colic. Two individual-level variables of interest were the age of the horse, X, and an indicator, Z, of whether the horse had undergone a recent change in diet.

In this example it is difficult to model parametrically the "next sick horse in the clinic" matching method by explicit variables S_o. One can certainly try, for example, by accounting for month, veterinarian, region of Texas, urban/rural, another data, but the dimensionality is likely to be high, and clearly there is more to the matching than can be measured explicitly. However, it is entirely possible that some or all of these factors affect the distribution of X among the controls, and a likelihood analysis would have to account for them. The dimensionality of the effects quickly gets out of hand in our example, and it is likely

Samiran Sinha is Assistant Professor of Statistics, Texas A&M University, College Station, TX 77843 (E-mail: *sinha@stat.tamu.edu*). Bhramar Mukherjee is Assistant Professor (E-mail: *mukherjee@stat.ufl.edu*) and Malay Ghosh is Distinguished Professor of Statistics (E-mail: *ghoshm@stat.ufl.edu*), Department of Statistics, University of Florida, Gainesville, FL 32611. Bani K. Mallick is Professor (E-mail: *bmallick@stat.tamu.edu*) and Raymond J. Carroll is Distinguished Professor of Statistics (E-mail: *carroll@stat.tamu.edu*), Department of Statistics, Texas A&M University, College Station, TX 77843. The research of Sinha and Ghosh was supported in part by National Institutes of Health grant R01 85414. The research of Mukherjee was supported in part by a National Science Foundation (NSF) New Researcher's Fellowship sponsored through Stanford University. The research of Mallick and Carroll was supported in part by a grant from National Cancer Institute (CA-57030), and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106). Mallick's research was also supported in part by NSF grant DMS-02-03215. The authors wish to thank the editor, associate editor, the referees, and Peter Müller for their comments on earlier drafts of the paper.

Journal of the American Statistical Association
June 2005, Vol. 100, No. 470, Theory and Methods
DOI 10.1198/016214504000001411

that in such cases the possible effect of stratification on the distribution of X will be ignored.

A second general example is also of interest. There has been a recent resurgence of interest in genetic case-control studies to explore a variety of gene–disease and gene–environment interactions. The prime objective here is to examine the association between a candidate gene and the occurrence of a disease. In such problems, it is very common to stratify a population that comprises subpopulations with different allele frequencies for the gene under consideration. If different subpopulations have different risks for the disease, then failure to take these stratum effects into account will lead to misleading estimates of the association between the disease and the candidate gene. (See Satten, Flanders, and Yang 2001 for a nice description of this phenomenon in the unmatched case-control context.) In our case of a matched case-control study, population stratification is handled by the matching when there are no missing data, but interestingly, with missing-data, population stratification may become an issue. Our approach can be viewed as a semiparametric method for handling such an issue if the stratification is caused by a continuous variable.

The likelihood approaches referenced earlier start with an explicit parametric distribution for X among the controls, $D = 0$, and given $\mathbf{Z}$ as well as possibly the explicit parts S_o of the strata. Wang et al. (1997) and Paik and Sacco (2000) did not include the strata in the model for X, whereas Satten and Carroll (2000) included the observed components S_o as a linear term. In the equine example, we of course believe that it is effectively impossible to measure all of the stratification variables and then model their effect on X, which is likely to be in the form of an unexplained heterogeneity.

Our approach is to generalize the previous methods to allow for the possibility of unmeasured stratification effects, that is, unexplained heterogeneity in the distribution of X given $\mathbf{Z}$ among the controls. Alternatively, our approach can be looked at as a way of allowing for high-dimensional effects of stratification. We model the stratification effects via a Dirichlet process prior with a normal base measure for the distribution of the stratum effects, then use the Bayesian machinery. By this route, we achieve a measure of model robustness and acknowledge that the distribution of X among the controls can be affected by stratification, especially by unmeasured factors.

The outline of the article is as follows. Section 2 introduces the model and notation, and Section 3 derives the appropriate likelihoods and the Markov chain Monte Carlo (MCMC) method for computing Bayesian inference. We note in passing that a limiting case of our approach is a full parametric Bayesian analysis of matched case-control studies with missing data, something that does not appear to be available in the literature.

Section 4 provides data analysis for three examples covering two special cases of the general exponential family: (a) a continuous exposure in the equine epidemiology problem, (b) a binary exposure in an endometrial cancer study, and (c) a binary exposure in a low birth weight study in newborns. Section 5 contains a small simulation study to assess the performance of our proposed methods. We show that at least for this simulation, our methods lose nothing in the way of efficiency when there are no unmeasured stratification effects on the missing covariate X, but gain quite a bit of efficiency when stratification does affect the missing covariate. Section 6 presents some concluding remarks. All technical details are collected into the Appendix.

In concluding this section, we highlight some of the newer aspects of this article. To our knowledge, this is the first article that takes a Bayesian approach, both parametric and semiparametric, toward the analysis of matched case-control studies with missing exposure. Our method is applicable to both discrete and continuous data. Moreover, our proposed method explicitly captures the unmeasured stratum effects, which may not be possible just by introducing additional covariates.

2. MODEL AND NOTATION

For subject j in stratum i, $i = 1, \ldots, n$, $j = 1, \ldots, M + 1$, let D_{ij} represent the disease status, namely $D_{ij} = 1$ for a case and $D_{ij} = 0$ for a control. We consider a single exposure variable X_{ij}, which may be partially missing, and a completely observed p-dimensional vector of covariates $\mathbf{Z}_{ij}$. Throughout, we assume that the $\mathbf{Z}_{ij}$ are nonstochastic. Let $\mathbf{S}_i$ be the collection of measured and unmeasured stratification variables for stratum i. We assume that the prospective conditional logistic distribution for the disease status is

$$\Pr(D_{ij} = 1 | X_{ij}, \mathbf{Z}_{ij}, \mathbf{S}_i) = H\{\beta_0(\mathbf{S}_i) + \boldsymbol{\beta}_1^T \mathbf{Z}_{ij} + \beta_2 X_{ij}\}, \qquad (1)$$

where $H(u) = \{1 + \exp(-u)\}^{-1}$. Here $\beta_0(\mathbf{S}_i)$ is a term representing the stratum effect on the probability of belonging to a particular disease state, and $\boldsymbol{\beta}_1$ and β_2 are coefficients corresponding to the covariates and exposure in the foregoing logistic regression model. The usual method is to work with the conditional likelihood of the D_{ij} with $\sum_{j=1}^{M+1} D_{ij}$, $i = 1, \ldots, n$, as the conditioning variables. In this way, the nuisance parameters $\beta_0(\mathbf{S}_i)$ are eliminated.

In the presence of missing X_{ij}, a naive application of the foregoing method leads to removal of the corresponding subject from the analysis even though observations on $(D, \mathbf{Z})$ for the subject are observed. On the other hand, if a case has a missing observation on the variable X, then the foregoing method leads to deletion of the entire matched set to which the case belongs even though observations on $(D, \mathbf{Z}, X)$ for other subjects of the matched set may still be available. This naturally entails a loss of information. To overcome this problem, we adopt the following approach.

Let Δ_{ij} represent an indicator variable that takes the value 0 if the exposure value is missing for subject $j = 1, \ldots, M + 1$ in stratum $i = 1, \ldots, n$ and 1 otherwise. The basic structure of the model is

$$\begin{aligned} p(D_{ij}, X_{ij}, \Delta_{ij} | \mathbf{Z}_{ij}, \mathbf{S}_i) &\\ = p(X_{ij} | D_{ij}, \Delta_{ij}, \mathbf{Z}_{ij}, \mathbf{S}_i) &\times p(\Delta_{ij} | D_{ij}, \mathbf{Z}_{ij}, \mathbf{S}_i) \\ &\times p(D_{ij} | \mathbf{Z}_{ij}, \mathbf{S}_i). \qquad (2) \end{aligned}$$

Following Satten and Carroll (2000), we assume that the $\mathbf{X}$'s are missing at random in the sense of Little and Rubin (1987), that is, $p(X_{ij} | D_{ij}, \Delta_{ij}, \mathbf{Z}_{ij}, \mathbf{S}_i) = p(X_{ij} | D_{ij}, \mathbf{Z}_{ij}, \mathbf{S}_i)$, and that $p(\Delta_{ij} | D_{ij}, \mathbf{Z}_{ij}, \mathbf{S}_i)$ does not depend on any parameters of interest.

Next, we model the distributions of the exposure variable X_{ij} among the controls as

$$p(X_{ij} | D_{ij} = 0, \mathbf{Z}_{ij}, \mathbf{S}_i) = \exp\left[\xi_{ij}\{\theta_{ij} X_{ij} - b(\theta_{ij})\} + c(X_{ij}, \xi_{ij})\right], \qquad (3)$$

where θ_{ij} is the natural parameter and is modeled as

$$\theta_{ij} = \gamma_{0i} + \boldsymbol{\gamma}^T \mathbf{Z}_{ij}, \tag{4}$$

where γ_{0i}, the varying intercept, captures the stratum effect on the natural parameter θ_{ij} and consequently the stratum effect on the exposure distribution. Paik and Sacco (2000) adopted a simpler parametric model with $\gamma_{0i} = \gamma_0$. If the stratification variables are written as $\mathbf{S}_i = (S_{oi}, S_{ui})$ with S_{oi} observed and S_{ui} unobserved, then Satten and Carroll (2000) set γ_{0i} in (4) to be a parametric linear function of S_{oi}. As argued in Section 1, it may not be possible to model the γ_{0i} explicitly in this way as a function of the stratification effects, and our methods differ from others in the fact that we address this issue.

In the next section we derive the joint distribution for (D_{ij}, X_{ij}) conditional on the sums $\sum_{j=1}^{M+1} D_{ij}$ and using (1–3).

3. LIKELIHOOD, PRIOR, AND POSTERIOR

In this section we derive the likelihood function, state our prior distribution, and derive the posterior. The key aspect of the modeling is in how we handle the strata effects in the model for the missing covariate among the controls.

In what follows we do calculations concerning X as if it were a continuous variable and hence use integration, if X is discrete, then the integrals are replaced by sums. The key result is establishing the odds when X is not considered. Using calculations similar to those of Satten and Carroll (2000), we show in the Appendix that (1) and (3) together imply that

$$\frac{\Pr(D_{ij}=1|\mathbf{Z}_{ij}, \mathbf{S}_i)}{\Pr(D_{ij}=0|\mathbf{Z}_{ij}, \mathbf{S}_i)} = \exp\bigl[\beta_0(\mathbf{S}_i) + \boldsymbol{\beta}_1^T \mathbf{Z}_{ij} + \xi_{ij}\{b(\theta_{ij}^*) - b(\theta_{ij})\}\bigr], \tag{5}$$

where $\theta_{ij}^* = \theta_{ij} + \xi_{ij}^{-1}\beta_2$. Result (5) is important because it means that, prospectively, the marginal (over X) probability of an event is logistic, and hence standard conditional logistic regression calculations can be used.

We also show in the Appendix that

$$p(X_{ij}|D_{ij}=1, \mathbf{Z}_{ij}, \mathbf{S}_i) = \exp\bigl[\xi_{ij}\{\theta_{ij}^* X_{ij} - b(\theta_{ij}^*)\} + c(X_{ij}, \xi_{ij})\bigr]. \tag{6}$$

As noted by a referee, there is an underlying asymmetry in (5) and (6) in treating the partially missing covariate X and the completely observed covariate $\mathbf{Z}$, with the latter assumed to be nonstochastic. In principle one could also model the completely observed covariate $\mathbf{Z}$ in a retrospective fashion by assuming a parametric distribution for $p(\mathbf{Z}_{ij}|D_{ij}=0, \mathbf{S}_i)$. Adopting that model will lead to

$$\frac{\Pr(D_{ij}=1|\mathbf{S}_i)}{\Pr(D_{ij}=0|\mathbf{S}_i)} = \int \frac{\Pr(D_{ij}=1|\mathbf{Z}_{ij}, \mathbf{S}_i)}{\Pr(D_{ij}=0|\mathbf{Z}_{ij}, \mathbf{S}_i)} \times p(\mathbf{Z}_{ij}|D_{ij}=0, \mathbf{S}_i)\, d\mathbf{Z}_{ij}.$$

The difficulty with this approach is in terms of model robustness. The covariates $\mathbf{Z}$ are usually multivariate with a mixture of categorical and continuous covariates, so that specifying an accurate model for its distribution may not be an easy task. Indeed, if the specified model for $p(\mathbf{Z}_{ij}|D_{ij}=0, \mathbf{S}_i)$ is incorrect, then we risk the possibility of biased and inconsistent estimation.

Instead of working with the full likelihood, we use a conditional likelihood, where the conditioning variables are $\sum_{j=1}^{M+1} D_{ij}$, the number of cases in the matched set i ($i = 1, 2, \ldots, n$). Unlike in an unmatched case-control study, in a matched study the total number of cases in each matched set is necessarily prespecified by the study design. This outcome dependent sampling structure is built into the conditional likelihood. Indeed, the conditioning is not merely an artificial device to eliminate the nuisance parameters $\beta_0(\mathbf{S}_i)$, but is forced by the matched design.

Breslow, Day, Halvorsen, Prentice, and Sabai (1978) compared the conditional and unconditional likelihood methods for the Singapore esophageal cancer study and noted the problems of unconditional likelihood inference with large number of stratum-specific parameters (see Breslow 1996, pp. 20–21, for details). In a Bayesian framework one can always model the stratum effect parameters $\beta_0(\mathbf{S}_i)$ in an unconditional likelihood (without conditioning on $\sum_{j=1}^{M+1} D_{ij}$), but then any comparison with the CLR or Bayesian analog of the model of Satten and Carroll (2000), as done in this article, seems infeasible.

We assume, without loss of generality, that $D_{i1} = 1$, $D_{i2} = 0, \ldots, D_{iM+1} = 0$, for each stratum i. Note that we need only define a probability model for $[D_{ij}|X_{ij}, \mathbf{Z}_{ij}, \mathbf{S}_i]$ and $[X_{ij}|D_{ij}=0, \mathbf{Z}_{ij}, \mathbf{S}_i]$ to arrive at the joint conditional likelihood of $[D_{ij}, X_{ij}|\mathbf{Z}_{ij}, \mathbf{S}_i, \sum_{j=1}^{M} D_{ij} = 1, \Delta_{ij}]$. The expression for this likelihood is based on the following key factorization:

$$\begin{aligned}
L_c &\propto \prod_{i=1}^{n} \Pr\Biggl\{(D_{i1}=1, D_{ij}=0, j \neq 1), \\
&\qquad\qquad \{X_{ij}, \Delta_{ij}\}_{j=1}^{M+1} \Bigm| \mathbf{S}_i, \mathbf{Z}_{ij}, \sum_{j=1}^{M+1} D_{ij} = 1\Biggr\} \\
&\propto \prod_{i=1}^{n} \Pr\Biggl\{D_{i1}=1, D_{ij}=0, j \neq 1 \Bigm| \mathbf{S}_i, \{\mathbf{Z}_{ij}\}_{j=1}^{M+1}, \sum_{j=1}^{M+1} D_{ij} = 1\Biggr\} \\
&\qquad \times p\bigl(\{X_{ij}\}_{j=1}^{M+1} \bigm| \mathbf{S}_i, \{\mathbf{Z}_{ij}, \Delta_{ij}\}_{j=1}^{M+1}, D_{i1}=1, D_{ij}=0, j \neq 1\bigr) \\
&= \prod_{i=1}^{n} \Biggl\{p^{\Delta_{i1}}(X_{i1}|\mathbf{S}_i, \mathbf{Z}_{i1}, D_{i1}=1) \\
&\qquad\qquad \times \prod_{j=2}^{M+1} p^{\Delta_{ij}}(X_{ij}|\mathbf{S}_i, \mathbf{Z}_{ij}, D_{ij}=0)\Biggr\} \\
&\qquad \times \prod_{i=1}^{n} \frac{\Pr(D_{i1}=1|\mathbf{S}_i, \mathbf{Z}_{i1}) \prod_{j=2}^{M+1} \Pr(D_{ij}=0|\mathbf{S}_i, \mathbf{Z}_{ij})}{\sum_{l=1}^{M+1} \Pr(D_{il}=1|\mathbf{S}_i, \mathbf{Z}_{il}) \prod_{j \neq l}^{M+1} \Pr(D_{ij}=0|\mathbf{S}_i, \mathbf{Z}_{ij})} \\
&= \prod_{i=1}^{n} \Biggl\{p^{\Delta_{i1}}(X_{i1}|\mathbf{S}_i, \mathbf{Z}_{i1}, D_{i1}=1) \\
&\qquad\qquad \times \prod_{j=2}^{M+1} p^{\Delta_{ij}}(X_{ij}|\mathbf{S}_i, \mathbf{Z}_{ij}, D_{ij}=0)\Biggr\} \\
&\qquad \times \prod_{i=1}^{n} \frac{\Pr(D_{i1}=1|\mathbf{S}_i, \mathbf{Z}_{i1}) / \Pr(D_{i1}=0|\mathbf{S}_i, \mathbf{Z}_{i1})}{\sum_{l=1}^{M+1} \mathrm{pr}(D_{il}=1|\mathbf{S}_i, \mathbf{Z}_{il}) / \Pr(D_{il}=0|\mathbf{S}_i, \mathbf{Z}_{il})}.
\end{aligned} \tag{7}$$

Using (5)–(7), the conditional likelihood is given by

$$\begin{aligned}
L_c(\boldsymbol{\beta}_1, \beta_2, \boldsymbol{\gamma}, \gamma_{01}, \gamma_{02}, \dots, \gamma_{0n})
&\propto \left\{ \prod_{i=1}^{n} p^{\Delta_{i1}}(X_{i1}|\mathbf{Z}_{i1}, \mathbf{S}_i, D_{i1} = 1) \right. \\
&\qquad \left. \times \prod_{j=2}^{M+1} p^{\Delta_{ij}}(X_{ij}|\mathbf{Z}_{ij}, \mathbf{S}_i, D_{ij} = 0) \right\} \\
&\qquad \times \frac{\exp[\sum_{i=1}^{n} \boldsymbol{\beta}_1^T \mathbf{Z}_{i1} + \sum_{i=1}^{n} \xi_{i1}\{(b(\theta_{i1}^*) - b(\theta_{i1})\}]}{\prod_{i=1}^{n}[\sum_{j=1}^{M+1} \exp[\boldsymbol{\beta}_1^T \mathbf{Z}_{ij} + \xi_{ij}\{b(\theta_{ij}^*) - b(\theta_{ij})\}]} \\
&= \prod_{i=1}^{n} \exp\big[\xi_{i1}\Delta_{i1}\{\theta_{i1}^* X_{i1} - b(\theta_{i1}^*)\} + \Delta_{i1} c(X_{i1}, \xi_{i1})\big] \\
&\qquad \times \prod_{i=1}^{n} \prod_{j=2}^{M+1} \exp\big[\Delta_{ij}\xi_{ij}\{\theta_{ij} X_{ij} - b(\theta_{ij})\} + \Delta_{ij} c(X_{ij}, \xi_{ij})\big] \\
&\qquad \times \prod_{i=1}^{n} \left(1 + \sum_{j=2}^{M+1} \exp\big[\boldsymbol{\beta}_1^T(\mathbf{Z}_{ij} - \mathbf{Z}_{i1}) + \xi_{ij}\{b(\theta_{ij}^*) - b(\theta_{ij})\} \right. \\
&\qquad\qquad \left. - \xi_{i1}\{b(\theta_{i1}^*) - b(\theta_{i1})\}\big] \right)^{-1}. \qquad (8)
\end{aligned}$$

The main problem in (8) is to estimate the regression parameters $\boldsymbol{\beta}_1$ and β_2. However, it may be noted that although the foregoing conditional likelihood has eliminated the nuisance parameters $\beta_0(\mathbf{S}_i)$, the nuisance parameters γ_{0i} remain (via θ_{ij}), and they increase in direct proportion to the number of strata. Hence even the conditional maximum likelihood of $\boldsymbol{\beta}_1$ or β_2 is potentially subject to the Neyman–Scott phenomenon and produces inconsistent estimators. Adopting a Bayesian approach is one way to circumvent this difficulty.

We consider the following set of mutually independent priors, $\boldsymbol{\beta}_1 \sim \text{normal}(\boldsymbol{\mu}_{\boldsymbol{\beta}_1}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_1})$, $\beta_2 \sim \text{normal}(\mu_{\beta_2}, \sigma_{\beta_2}^2)$, $\gamma_j \overset{\text{iid}}{\sim} \text{normal}(\mu_*, \sigma_*^2)$, and $\gamma_{0i}|G \overset{\text{iid}}{\sim} G$, where G has a Dirichlet process prior with a normal base measure. Symbolically, $G \sim \text{DP}(\alpha G_0)$ and G_0 is $\text{N}(\zeta_0, \sigma_0^2)$, where α is the concentration parameter. Using the result of Antoniak (1974), it follows that the predictive distribution of $\gamma_{0\overline{n+1}}$ given $\gamma_{01}, \dots, \gamma_{0n}$ is

$$\gamma_{0\overline{n+1}}|\gamma_{01}, \dots, \gamma_{0n} \sim \frac{\alpha}{\alpha + n} G_0(\cdot) + \frac{1}{\alpha + n} \sum_{k=1}^{n} I_{\gamma_{0k}}(\cdot). \qquad (9)$$

Remark 1. Note that for very large values of α relative to n, the Bayesian semiparametric (BSP) method produces essentially the $\text{N}(\zeta_0, \sigma_0^2)$ predictive distribution for $\gamma_{0\overline{n+1}}$, whereas very small values of α relative to n amounts to the fact that the predictive distribution of $\gamma_{0\overline{n+1}}$ essentially equals the "empirical" distribution of $\gamma_{01}, \dots, \gamma_{0n}$. We also note that as $\alpha \to \infty$, the Dirichlet process model reduces to specifying a parametric model on the γ_{0i}, namely $\gamma_{0i} \overset{\text{iid}}{\sim} G_0$, whereas $\alpha = 0$ simply implies a parametric model with a constant stratum effect, namely $\gamma_{0i} = \gamma_0$ and $\gamma_0 \sim G_0$.

In our computations we assume a gamma prior on α and resample from the full conditional distribution of α using a latent beta variable, as described by Escobar and West (1995).

With the foregoing model and prior specifications, one can obtain the full conditional distributions for the parameters $\boldsymbol{\beta}_1, \beta_2$, and $\gamma_j, j = 1, \dots, p$ and $\gamma_{0i}, i = 1, \dots, n$. The exact expressions are given in the Appendix. For estimation of parameters, we use a MCMC computing scheme. The simulation strategy is discussed in detail in the Appendix.

Remark 2. It is easy to extend the results of this section to $\mathbf{X}_{ij} = (X_{ij1}, \dots, X_{ijq})^T$, where $\mathbf{X}_{ij}$ has a distribution belonging to the multiparameter exponential family. We then have to introduce the indicator variables $\Delta_{ijk} = 1$ or 0 according to whether X_{ijk} is observed or missing. The likelihood given in (8) then needs to be suitably modified. Modeling association among multiple exposures with joint distribution outside the multiparameter exponential family is an ongoing project.

4. THREE SPECIAL CASES WITH DATA EXAMPLES

It is worth noting that the proposed model can accommodate both continuous and discrete exposure variables and also can model for the missingness in the exposure variable. In this section, we present three particular illustrations, one with a normally distributed exposure variable and the other two with a binary exposure variable, all motivated by real examples.

4.1 A Continuous Exposure: The Equine Epidemiology Example

The first example is the equine epidemiology example described in Section 1. Age was considered as a single exposure variable (X) measured on a continuous scale with one binary covariate (Z) indicating whether the horse experienced recent diet changes or not. This is a 1:1 matched study with 498 strata. For scaling purposes, we linearly transformed age so that X was on the interval [0, 1]. Because we are dealing with a single covariate and one exposure variable here, the corresponding regression parameters are all scalar.

The proposed method for the general exponential family, as described earlier, applies here with continuous exposure X. We assume that conditional on $D_{ij} = 0$ (i.e., for the control population), the exposure variable age has a normal distribution of the following form:

$$[X_{ij}|D_{ij} = 0, Z_{ij}, \mathbf{S}_i] \sim \text{normal}(\gamma_{0i} + \gamma_1 Z_{ij}, \sigma^2), \qquad (10)$$

where γ_{0i} are stratum effects on the exposure distribution and γ_1 is the regression parameter for regressing X_{ij} on the measured covariates Z_{ij}. We also assume the logistic regression model for disease status as in (1). From (6), it follows that the exposure distribution among the cases is normal with mean $\gamma_{0i} + \gamma_1 Z_{ij} + \sigma^2\beta_2$ and variance σ^2. The conditional likelihood analogous to (8) can now be derived using these facts. We used the following prior distributions. The parameters $(\beta_1, \beta_2, \gamma_1)$ were independent normal(0, 10), and the γ_{0i}'s had a Dirichlet prior with base measure $G_0 \sim \text{normal}(0, 10)$. Experimentation suggested that the parameter estimates remain relatively unchanged over a range of gamma priors on α. Table 1 gives the results corresponding to a gamma(2, 2) prior on α. The variance σ^2 in (10) is assigned an inverse gamma prior, that is, $\pi(\sigma^2) \propto (\sigma^2)^{-(c/2)-1} \exp(-\frac{d}{2\sigma^2})$, written symbolically as $\text{IG}(c/2, d/2)$; for our example, $c = 60$ and $d = 2$.

The full conditional distributions for all of the parameters have compact expressions that can be derived as special cases

Table 1. Results for the Equine Data Example

Method		Full equine data β_1	β_2	$10\times\gamma_1$	$10\times\sigma^2$	Equine data with 40% missing data β_1	β_2	$10\times\gamma_1$	$10\times\sigma^2$
BSP									
	Mean	2.16	2.18	.18	.23	2.16	2.30	.14	.23
	SD	.33	.39	.14	.01	.32	.48	.24	.01
	Lower HPD	1.57	1.36	−.10	.21	1.70	1.30	−.25	.19
	Upper HPD	2.80	2.88	.40	.25	2.80	3.42	.60	.25
PB									
	Mean	2.14	2.10	.22	.26	2.13	1.90	.17	.28
	SD	.32	.41	.17	.01	.32	.47	.20	.02
	Lower HPD	1.60	1.32	−.20	.24	1.56	.92	−.20	.24
	Upper HPD	2.87	2.88	.60	.28	2.85	2.85	.67	.31
CLR									
	Mean	2.13	2.05			2.81	2.79		
	SD	.32	.47			.59	.77		
	Lower CL	1.50	1.13			1.65	1.47		
	Upper CL	2.76	2.97			3.96	4.49		

NOTE: Here "mean" is the posterior mean, "SD" is the posterior standard deviation, "lower HPD" and "upper HPD" are the lower and upper ends of the 95% HPD region, "lower CL" and "upper CL" are the lower and upper ends of the confidence limit. The parametric Bayesian method is the Bayesian version of the method of Satten and Carroll (2000).

of equations (A.1)–(A.5) in the Appendix. The full conditional distribution for the additional scale parameter σ^2 is given by

$$\pi(\sigma^2|\cdot) \propto \sqrt{\frac{\lambda}{2\pi}}\,\frac{1}{(\sigma^2)^{3/2}} \times \exp\left\{-\frac{\lambda}{2\mu^2}\frac{(\sigma^2-\mu)^2}{\sigma^2}\right\}(\sigma^2)^{-(2n+c-1)/2}, \qquad (11)$$

where $\lambda = -2(a - d/2)$, $\mu = \{(a - d/2)/b\}^{1/2}$, $a = -\frac{1}{2}\sum_{i=1}^{n}\{(X_{i1}-\theta_{i1})^2 + (X_{i2}-\theta_{i2})^2\}$, and $b = -(n/2)\beta_2^2$.

For estimation of the parameters, we adopt a MCMC numerical integration scheme. For simulating observations from the posterior distribution of γ_{0i}, we follow the idea of Escobar (1994), West, Müller, and Escobar (1994), Escobar and West (1995), and Neal (2000). For generating random numbers from the conditional distribution of other parameters we follow a standard componentwise Metropolis–Hastings scheme (Robert and Casella 1999). The details of the computing scheme are provided in the Appendix.

For each of our examples and simulated datasets, we conduct three analyses. One is our proposed BSP analysis. The other is a parametric Bayesian (PB) version of the work of Satten and Carroll (2000) by considering a constant stratum effect $\gamma_{0i} \equiv \gamma_0$ and then assuming a normal prior on this common stratum effect parameter γ_0. We emphasize here and in the tables that the PB method is simply the Bayesian version of the method of Satten and Carroll (2000). We also present the frequentist alternative for analyzing this matched data through a standard CLR analysis (Breslow and Day 1980).

We ran these three analyses on the equine dataset with all 498 strata. We also reran the analyses where we randomly deleted 40% of the age observations. The results are given in Table 1.

Briefly, with no missing data, the results of the methods are roughly in accordance with one another. The slightly small standard error estimates for the Bayesian methods may be a reflection of the fact that unlike CLR, they are full likelihood-based semiparametric and parametric methods. With missing data, of course, we see that the Bayesian methods yield much smaller standard errors than CLR, reflecting the fact that our methods use all of the data, not just the pairs with no missing data. There is little difference between the BSP analysis and the PB version of the method of Satten and Carroll (2000), a phenomenon that can be explained as follows.

Figure 1 shows a boxplot of the variance of γ_{0i}'s across the 498 strata in the last 5,000 MCMC samples. The average variance among the γ_{0i}'s is very small (about .0054), suggesting that the stratum effects are almost constant across strata. Thus the model with constant stratum effect essentially holds for this example, and hence it is not surprising that the two methods

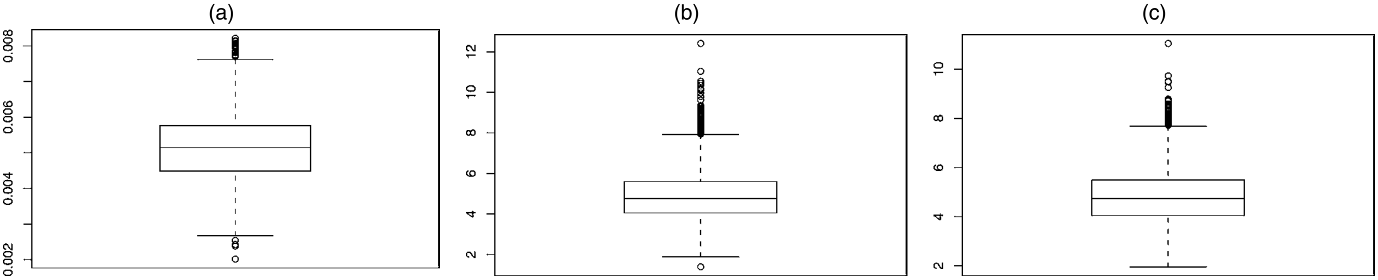

Figure 1. Boxplot of the Variability of γ_{0i}'s for All Three Examples: (a) Equine Data; (b) LA Cancer Data; (c) Low Birth Weight Data. For each example, estimates of the γ_{0i}'s (i = 1, . . . , n) were collected for each of the last 5,000 MCMC samples. Variance of these values were then calculated for each run. Each boxplot is based on these 5,000 variance values.

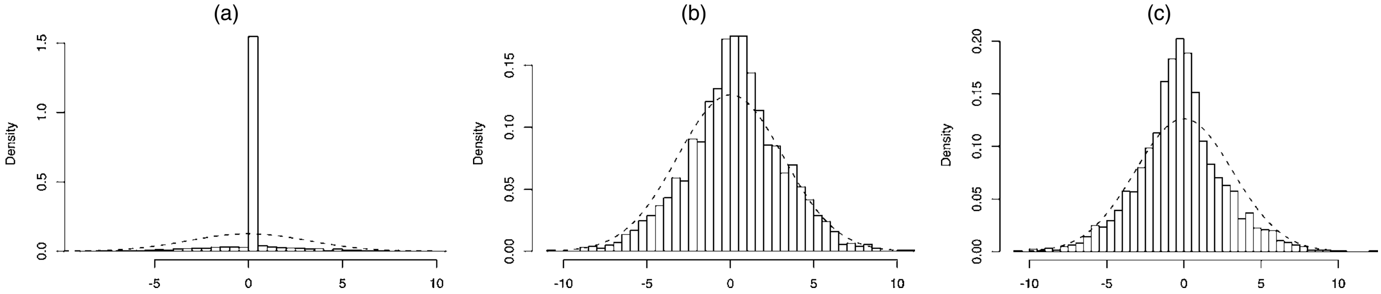

Figure 2. Posterior Predictive Density of $\gamma_{0\overline{n+1}}$ for the Three Examples: (a) Equine Data; (b) LA Cancer Data; (c) Low Birth Weight Data. $\gamma_{0\overline{n+1}}$ values are generated from the corresponding predictive distribution [as given in (9)] for each of the last 5,000 MCMC runs. The histogram is based on these 5,000 generated values. The dashed line corresponds to the density of the base measure normal(0, 10).

give similar results. Note that under our Dirichlet process model ($\gamma_{0i}|G \overset{\text{iid}}{\sim} G$, $i = 1, \dots, n$, with $G \sim \text{DP}(\alpha G_0)$), $E(G|Data)$ is in fact the posterior predictive distribution $p(\gamma_{0\overline{n+1}}|Data)$. Figure 2 presents a histogram for the predictive density of $\gamma_{0\overline{n+1}}$ given the data. For each of the last 5,000 MCMC runs, we generated $\gamma_{0\overline{n+1}}$ from the corresponding predictive distribution [as stated in (9)]. The histogram is drawn based on these 5,000 generations of $\gamma_{0\overline{n+1}}$ values. It is evident from this figure that the distribution is different from a simple normal model. The posterior distribution of the constant stratum effect γ_0 for the PB model is given in Figure 3.

Posterior inference on the concentration parameter α for all three examples is contained in Table 6. Note that for the equine data, the posterior mean of α (144.29) is small relative to n, the total number of stratum specific parameters (498). Consequently, the average number of distinct γ_{0i} values (38) is relatively small compared with n (498 for this data). This is consistent with the small variability in the γ_{0i} values as depicted in Figure 1.

4.2 A Binary Exposure: Endometrial Cancer

In the Los Angeles 1:4 matched case-control study on endometrial cancer (Breslow and Day 1980), the variable obesity may be considered a binary exposure variable (X) for contracting endometrial cancer. This variable had about 16% missing observations. We considered the presence of gall bladder disease in the subject as our completely observed dichotomous covariate Z.

In such a case with a dichotomous exposure variable, the exposure distribution for the control population is naturally assumed to have a logistic form: $\Pr(X_{ij} = 1|D_{ij} = 0, Z_{ij}, \mathbf{S}_i) = H(\gamma_{0i} + \gamma_1 Z_{ij})$, where $H(\cdot)$ is the logistic distribution function. From (6), the exposure distribution in the case population is again of the logistic form with $\Pr(X_{ij} = 1|D_{ij} = 1, Z_{ij}, \mathbf{S}_i) = H(\gamma_{0i} + \gamma_1 Z_{ij} + \beta_2)$.

In the LA study there were 63 strata. For this example also we adopted the same Metropolis–Hastings scheme as that of Neal (2000) for simulating observations from the posterior of γ_{0i}'s. Alternately, one can use the approach of MacEachern and Müller (1998), but Neal's approach seems easier to implement in our case. A detailed outline of the algorithm and computing scheme is contained in the Appendix. We present the results with a gamma$(2, 2)$ prior on α. We considered a normal$(0, 10)$ prior on all of the parameters as well as the base measure in the Dirichlet process prior.

The original data contained missing observations on the exposure variable obesity. We modeled the missingness in the exposure variable in both the PB (the Bayesian version of the Satten–Carroll method with fixed stratum effects) and the BSP analysis. We also conducted a classical conditional logistic analysis. Table 2 presents the results. Although there are certainly some interesting numerical differences, in the main the results are reasonably comparable.

Figure 1 presents a boxplot of the variance of the 63 stratum effects in the last 5,000 MCMC samples with the BSP method. The average variance among the γ_{0i}'s is approximately 4.56. This is fairly large, and if there were major covariate effects, then we would expect the BSP and the PB version of the method

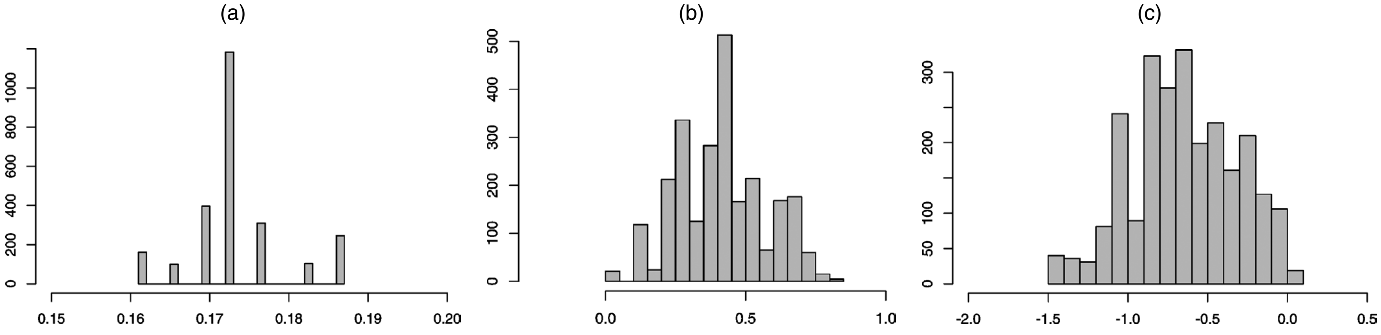

Figure 3. Histograms of the Posterior Distribution of γ_0 for the PB Analysis for Each of the Three Examples: (a) Equine Data; (b) LA Cancer Data; (c) Low Birth Weight Data.

Table 2. Results for the Endometrial Cancer Data Example

		Gall bladder obesity		
Method		β_1	β_2	γ_1
BSP				
	Mean	1.29	.70	−.15
	SE	.39	.41	.49
	Lower HPD	.54	−.10	−.99
	Upper HPD	2.18	1.49	.99
PB				
	Mean	1.19	.49	.42
	SE	.39	.36	.62
	Lower HPD	.58	−.15	−.70
	Upper HPD	2.12	1.25	1.78
CLR				
	Mean	1.28	.44	
	SE	.39	.38	
	Lower CL	.52	−.30	
	Upper CL	2.04	1.19	

NOTE: Here "mean" is the posterior mean, "SE" is the posterior standard deviation, "lower HPD" and "upper HPD" are the lower and upper end of the HPD region. "Lower CL" and "upper CL" are the lower and upper end of the confidence limit.

of Satten and Carroll (2000) to be different. However, in this example the analysis suggests that $\beta_2 = 0$ is plausible, so that D and X are essentially independent and whether or not the γ_{0i}'s are constant does not cause a model bias. This is an example in which there is stratum variability and natural missingness in the data. Figure 2 presents a plot of the posterior predictive density of $\gamma_{0\overline{n+1}}$.

Table 3 shows that the posterior mean of α (119.83) is large compared to n (63), and consequently the average number of distinct γ_{0i} values is large (48) compared with n. This is again consistent with the finding in Figure 1 that the average variability in the stratum effects for the LA cancer data is fairly large.

4.3 Another Binary Exposure: Low Birth Weight Study

In the equine data there was no appreciable stratum effect variability, although exposure had a significant effect on disease probability. The endometrial cancer data, on the other hand, has natural missingness and appreciable stratum variability, but the exposure variable obesity did not have a significant effect on disease probability. The example considered in this section illustrates a case where there is significant stratum variability and modeling it causes some interesting differences in inference for the exposure effect when compared with a constant stratum effect model.

This example involves a matched case-control dataset coming from a low birth weight study conducted by the Baystate Medical Center in Springfield, Massachusetts. This dataset was discussed by Hosmer and Lemeshow (2000, sec. 1.6.2) and was used as an illustrative example of analyzing a matched case-control study in chapter 7 of their book. Low birth weight (defined as birth weight $<2{,}500$ grams) is a cause of concern, because infant mortality and birth defect rates are very high for low-birth-weight babies. The data were matched according to the age of the mother. A woman's behavior during pregnancy (i.e., smoking habits, diet, prenatal care) can greatly alter the likelihood of carrying the baby to term. The goal of the study was to determine whether these variables were risk factors in the clinical population served by Baystate Medical Center. The matched data contain 29 strata, and each stratum has 1 case (a low-birth-weight baby) and 3 controls (normal-birth-weight babies). One can possibly think of many different models for explaining the disease in terms of the possible covariates recorded in the dataset. We consider smoking status of the mother as a single exposure variable. Two other covariates, a binary variable denoting presence of uterine irritability (UI) in the mother and weight of the mother at the last menstrual period (LWT), are also included in the model.

*Table 3. Posterior Inference on α and k**

		α		k	
Example	*n*	*Mean*	*SD*	*Mean*	*SD*
Equine data	498	144.29	30.07	38	5.17
LA cancer data	63	119.83	18.04	48	3.70
Low birth weight data	29	48.73	12.40	18	2.47

* k is the number of distinct components of $\{\gamma_{0i}\}_{i=1}^{n}$.

Tables 4 and 5 give the results of our analysis, the former for the full data and the latter when we induce 30% missingness in smoking status. Note that the regression coefficient for smoking status changes hardly at all in the semiparametric case when we induce missing data, but there are fairly substantial changes for CLR and the PB version of Satten and Carroll, which assume a constant stratum effect. Moreover, in the full data case, the effect of smoking is more pronounced with varying stratum effects than with constant stratum effects. Indeed, the strong association between smoking and low birth weight has been established (Walsh 1994).

Figure 1 shows that the average estimated stratum variability is about 4.43 for this example, justifying the need for a varying stratum effect model even in the absence of any missing data. Figure 2 provides the posterior predictive distribution of $\gamma_{0\overline{n+1}}$.

Table 3 shows that the posterior mean of α (48.73) is again high relative to n (29) and that the average number of distinct γ_{0i} values is large (18) compared with the total number of stratum-specific parameters (29 in this case). This finding is in agreement with Figure 1, showing significant variability among the γ_{0i} values for this example.

5. SIMULATION STUDY

To simulate a realistic dataset for comparing the BSP method with the PB version of the method of Satten and Carroll (2000) and standard matched CLR, we used the equine data itself as a prototype, except that we reversed the roles of diet change and age. We generated a hypothetical 1:1 matched dataset with 50 strata and with diet change (a binary variable) as the exposure (X) and age of horse (transformed on to $[0,1]$) as an observed covariate (Z). To elicit a realistic value for the true parameters $(\beta_1, \beta_2, \gamma_1)$, we made use of the original equine dataset at hand. We first performed a conditional logistic regression analysis of the full equine dataset with X and Z as covariates. The parameter estimates for the coefficients of X and Z using the CLR analysis were chosen as the true values of β_1 and β_2 in our simulation; these were 2.049 and 2.129. To elicit values for γ_1, we ran a logistic regression with diet change X as the response and age Z as the covariate. The fitted equation had logits $-3.66 + .63Z$. In the simulation study we thus used $\gamma_1 = .63$. The purpose of our simulation was to explore the relative performance of the three methods when in fact $\gamma_{0i} \equiv \gamma_0$,

Table 4. Full Data Analysis of Low Birth Weight Data

	BSP			PB			CLR	
Parameter	*Mean*	*SD*	*HPD region*	*Mean*	*SD*	*HPD region*	*Estimate*	*SD*
SMOKE	1.21	.56	(.15, 2.31)	.96	.46	(−.01, 1.85)	.86	.45
LWT	−1.17	1.26	(−3.39, 1.78)	−1.34	1.25	(−3.75, 1.15)	−1.13	1.36
UI	.91	.49	(.03, 1.98)	.79	.52	(−.31, 1.86)	.85	.51

NOTE: PB stands for parametric Bayes method assuming constant stratum effects. Here "mean" is the posterior mean, "SD" is the posterior standard deviation, "HPD region" is the 95% highest posterior density credible region.

the standard model assumptions were true, and also for the situation when the constant stratum effect assumption does not hold and γ_{0i} came from a distribution with reasonable variability.

We followed the structure of our models as described in Section 2 for simulating X, D, and Z. We started by generating the covariate Z. For the equine data, the histogram of the variable horse age resembled a distribution that could be well described by a gamma distribution with shape parameter 1.8 and scale parameter .2. Accordingly, we generated our covariate Z from the same gamma distribution and then scaled the generated values to $[0, 1]$. The results remain similar if we generate Z from a normal distribution.

Second, we generated the binary variable D standing for disease status. For the ith stratum, it may easily be noted that

$$\Pr(D_{i1} = 1 | D_{i1} + D_{i2} = 1, Z_{i1}, Z_{i2}, \mathbf{S}_i) = 1 - \frac{Q_{i2}}{Q_{i2} + \exp[\beta_1 (Z_{i1} - Z_{i2})] Q_{i1}},$$

where, for $j = 1, 2,$

$$Q_{ij} = \frac{1 + \exp(\gamma_{0i} + \gamma_1 Z_{ij} + \beta_2)}{1 + \exp(\gamma_{0i} + \gamma_1 Z_{ij})}.$$

We generated a Bernoulli variable with the foregoing specified probability structure and conditional on the fact that the simulated value for D_{i1} is 1 (i.e., the subject is a member of the case population). We generated the binary exposure mimicking diet change (X) from a Bernoulli distribution with $\operatorname{logit}(p_i) = \gamma_{0i} + \gamma_1 Z_{i1} + \beta_2$; if the simulated value for D_{i1} is 0 (i.e., the subject is a member of the control population), then we generated X from a Bernoulli distribution with $\operatorname{logit}(p_i) = \gamma_{0i} + \gamma_1 Z_{i1}$. Once we simulated D_{i1} in the foregoing manner, we did so for the other observation in the ith stratum, $D_{i2} = 1 - D_{i1}$. We then simulated the corresponding exposure value X_{i2} from a Bernoulli distribution with either $\operatorname{logit}(p_i) = \gamma_{0i} + \gamma_1 Z_{i2}$ (when $D_{i2} = 0$) or $\operatorname{logit}(p_i) = \gamma_{0i} + \gamma_1 Z_{i2} + \beta_2$ (when $D_{i2} = 1$).

We performed two sets of simulations, one set with a constant value of γ_{0i}, namely −3.66 (this choice is described earlier), and the other with γ_{0i} simulated from a normal distribution with mean −3.66 and standard deviation 2.

In our simulations, the prior on γ_0 in the PB version of the method of Satten and Carroll (2000) and for the base measure in the Dirichlet process prior for the BSP method is normal$(0, 10)$. We used a gamma$(2, 2)$ prior on α. We replicated the simulations 50 times. For each replication, we also generated a dataset with 30% observations missing completely at random. The results are given in Table 6.

The results are fairly clear. In the case of fixed strata effects and no missing data, there are little in the way of differences among the BSP method, the PB counterpart to the frequentist method of Satten and Carroll, and ordinary CLR. However, if there is 30% missing data, then ordinary CLR is clearly less efficient. This might be considered a quantification of robustness of efficiency for the BSP method. On the other hand, if there are major unexplained stratum effects (i.e., the γ_{0i} have considerable variability), then we see that the methods separate, with the BSP method clearly dominating in terms of mean squared error (MSE).

6. CONCLUDING REMARKS

This article has considered matched case-control studies with partially missing exposure data. Parametric analysis requires handling stratum effects on the distribution of the exposure distribution. Previous methods either ignore such stratum effects or assume that they can be captured in their entirety by a parametric model.

We have presented the BSP approach for modeling the stratum effects. Our proposed method can handle both discrete and continuous exposures and also accounts for possible missingness in the exposure variable. Moreover, all three examples involving real data reveal that BSP adapts to the different levels of variability among the stratum effects, thereby inducing robustness of the procedure. Results based on simulated data indicate that in the presence of varying stratum effects, the BSP method outperforms the PB method and frequentist methods currently in the literature.

Table 5. Analysis of Low Birth Weight Data With 30% Missingness in the SMOKE Variable

	BSP			PB			CLR	
Parameter	*Mean*	*SD*	*HPD region*	*Mean*	*SD*	*HPD region*	*Estimate*	*SD*
SMOKE	1.19	.82	(−.45, 2.85)	.65	.59	(−.47, 1.98)	.65	.53
LWT	−1.12	1.27	(−3.51, 1.46)	−1.22	1.24	(−3.70, 1.21)	−1.06	1.62
UI	.96	.55	(−.01, 2.20)	.86	.53	(−.30, 1.91)	.88	.60

NOTE: PB stands for the parametric Bayes method assuming constant stratum effects. Here "mean" is the posterior mean, "SD" is the posterior standard deviation, "HPD region" is the 95% highest posterior density credible region.

Table 6. Results of the Simulation Study

Method		β_1	β_2	γ_1
Full data, fixed $\gamma_{0i} \equiv -3.66$				
BSP				
	Mean	1.9880	2.0949	.6938
	MSE	.7595	.5324	1.0342
PB				
	Mean	2.0089	1.8183	.6743
	MSE	.7785	.5159	.6454
CLR				
	Mean	2.3157	1.9496	
	MSE	2.0505	.4603	
Full data, varying $\gamma_{0i} = \text{normal}(-3.66, 4)$				
BSP	Mean	1.7838	2.2194	.4729
	MSE	.7126	.5042	.8015
PB				
	Mean	1.7812	1.4641	.4795
	MSE	.9550	.6488	.8333
CLR				
	Mean	2.6350	1.7747	
	MSE	2.2622	.5667	
30% missing data, fixed $\gamma_{0i} \equiv -3.66$				
BSP				
	Mean	2.0022	2.4606	.9617
	MSE	.7925	.8962	1.0962
PB				
	Mean	2.0141	1.8156	.6913
	MSE	.7905	.8561	.9125
CLR				
	Mean	2.9713	1.2689	
	MSE	5.6249	1.3182	
30% missing data, varying $\gamma_{0i} = \text{normal}(-3.66, 4)$				
BSP				
	Mean	1.7922	2.0150	.4935
	MSE	.8095	.8152	1.0444
PB				
	Mean	1.7535	1.5081	.5737
	MSE	.9690	.9444	1.3068
CLR				
	Mean	3.6817	1.2078	
	MSE	10.2676	1.3744	

NOTE: Here "mean" is the mean of the 50 estimates corresponding to the 50 simulated datasets. The true parameter values are $\beta_1 = 2.049$, $\beta_2 = 2.129$, and $\gamma_1 = .63$. The PB method is the Bayesian version of the method of Satten and Carroll (2000).

APPENDIX: PROOFS AND COMPUTATIONS

A.1 Proof of (5) and (6)

Note that

$$\begin{aligned} p(D_{ij} = 1|\mathbf{Z}_{ij}, \mathbf{S}_i) &= \int \exp\{\beta_0(\mathbf{S}_i) + \boldsymbol{\beta}_1^T \mathbf{Z}_{ij} + \beta_2 X_{ij}\} \\ &\qquad \times p(D_{ij} = 0|X_{ij}, \mathbf{Z}_{ij}, \mathbf{S}_i) p(X_{ij}|\mathbf{Z}_{ij}, \mathbf{S}_i)\, dX_{ij} \\ &= \exp\{\beta_0(\mathbf{S}_i) + \boldsymbol{\beta}_1^T \mathbf{Z}_{ij}\} \\ &\qquad \times \int \exp(\beta_2 X_{ij}) p(X_{ij}, D_{ij} = 0|\mathbf{Z}_{ij}, \mathbf{S}_i)\, dX_{ij} \\ &= \exp\{\beta_0(\mathbf{S}_i) + \boldsymbol{\beta}_1^T \mathbf{Z}_{ij}\} p(D_{ij} = 0|\mathbf{Z}_{ij}, \mathbf{S}_i) \\ &\qquad \times \int \exp(\beta_2 X_{ij}) p(X_{ij}|D_{ij} = 0, \mathbf{Z}_{ij}, \mathbf{S}_i)\, dX_{ij}. \end{aligned}$$

This proves (5). To prove (6), note that

$$\begin{aligned} p(X_{ij}|D_{ij} = 1, \mathbf{Z}_{ij}, \mathbf{S}_i) &= \frac{p(X_{ij}, D_{ij} = 1|\mathbf{Z}_{ij}, \mathbf{S}_i)}{\Pr(D_{ij} = 1|\mathbf{Z}_{ij}, \mathbf{S}_i)} \\ &= \frac{\Pr(D_{ij} = 1|X_{ij}, \mathbf{Z}_{ij}, \mathbf{S}_i) \times p(X_{ij}|\mathbf{Z}_{ij}, \mathbf{S}_i)}{\Pr(D_{ij} = 1|\mathbf{Z}_{ij}, \mathbf{S}_i)} \\ &= \frac{\exp(\beta_2 X_{ij}) \Pr(D_{ij} = 0, X_{ij}|\mathbf{Z}_{ij}, \mathbf{S}_i)}{\Pr(D_{ij} = 0|\mathbf{Z}_{ij}, \mathbf{S}_i) \exp[\xi_{ij}\{b(\theta_{ij} + \xi_{ij}^{-1}\beta_2) - b(\theta_{ij})\}]} \\ &= \frac{\exp(\beta_2 X_{ij}) \operatorname{pr}(D_{ij} = 0|\mathbf{Z}_{ij}, \mathbf{S}_i) p(X_{ij}|D_{ij} = 0, \mathbf{Z}_{ij}, \mathbf{S}_i)}{\Pr(D_{ij} = 0|\mathbf{Z}_{ij}, \mathbf{S}_i) \exp[\xi_{ij}\{b(\theta_{ij}^*) - b(\theta_{ij})\}]} \\ &= \exp[\xi_{ij}\{\theta_{ij}^* X_{ij} - b(\theta_{ij}^*)\} + c(X_{ij}, \xi_{ij})]. \end{aligned}$$

A.2 Full Conditional Distribution of the Parameters

For concise notation, let

$$g_{ij} = \exp\bigl[\boldsymbol{\beta}_1^T(\mathbf{Z}_{ij} - \mathbf{Z}_{i1}) + \xi_{ij}\{b(\theta_{ij}^*) - b(\theta_{ij})\} - \xi_{i1}\{b(\theta_{i1}^*) - b(\theta_{i1})\}\bigr]. \tag{A.1}$$

The full conditional distribution of the parameters may be obtained as

$$\pi(\boldsymbol{\beta}_1|\cdot) \propto \prod_{i=1}^{n}\left(1 + \sum_{j=2}^{M+1} g_{ij}\right)^{-1} \times \exp\left[-\frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_{\boldsymbol{\beta}_1})^T \boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_{\boldsymbol{\beta}_1})\right]; \tag{A.2}$$

$$\pi(\beta_2|\cdot) \propto \exp\left[\sum_{i=1}^{n} \Delta_{i1}\xi_{i1}\{\theta_{i1}^* X_{i1} - b(\theta_{i1}^*)\}\right] \times \prod_{i=1}^{n}\left(1 + \sum_{j=2}^{M+1} g_{ij}\right)^{-1} \times \exp\left[-\frac{1}{2\sigma_{\beta_2}^2}(\beta_2 - \mu_{\beta_2})^2\right]; \tag{A.3}$$

$$\begin{aligned} \pi(\gamma_j|\cdot) \propto{}& \exp\left[\sum_{i=1}^{n} \Delta_{i1}\xi_{i1}\{\theta_{i1}^* X_{i1} - b(\theta_{i1}^*)\}\right] \\ &\times \exp\left[\sum_{i=1}^{n}\sum_{j=2}^{M+1} \Delta_{ij}\xi_{ij}\{\theta_{ij} X_{ij} - b(\theta_{ij})\}\right] \\ &\times \prod_{i=1}^{n}\left(1 + \sum_{j=2}^{M+1} g_{ij}\right)^{-1} \\ &\times \exp\left[-\frac{1}{2\sigma_*^2}(\gamma_j - \mu_*)^2\right], \qquad j = 1, \ldots, p; \end{aligned} \tag{A.4}$$

and

$$\begin{aligned} \pi(\gamma_{0i}|\cdot) \propto{}& \frac{\alpha}{\alpha + n - 1} w_1(x_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}_1, \beta_2, \boldsymbol{\gamma}_1, \Delta_{ij}, \theta_{ij}, \xi_{ij}) \\ &\times \mathbf{N}\left(\zeta_0 + \sigma_0^2 \sum_{j=1}^{M+1} \Delta_{ij}\xi_{ij}x_{ij}, \sigma_0^2\right) \\ &+ \frac{1}{\alpha + n - 1} w_2(x_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}_1, \beta_2, \boldsymbol{\gamma}_1, \Delta_{ij}, \theta_{ij}, \xi_{ij}) \\ &\times \sum_{\substack{k=1\\k\neq i}}^{n} I_{\gamma_{0i}}(\gamma_{0k}), \end{aligned} \tag{A.5}$$

where I is the indicator function and w_1 and w_2 are two weight functions that can be calculated explicitly and are such that $\pi(\gamma_{0i}|\cdot)$ is a proper density. Thus expression (A.5) may be derived using the classical results of Antoniak (1974). We also note that θ_{ij} and θ_{ij}^* involve γ_j.

A.3 Computational Details

Note that none of the foregoing full conditionals has a standard distributional form, and thus simulating observations from the foregoing conditionals is not automatic. We adopted a componentwise Metropolis–Hastings algorithm for each of the parameters, namely $\boldsymbol{\beta}_1$, β_2, and $\boldsymbol{\gamma}$. For simulating observations from the posterior of γ_{0i}, we used an algorithm suggested by Neal (2000). The details of the computational steps are as follows.

For parameters other than γ_{0i}, $i = 1, \ldots, n$, we essentially followed the same Metropolis–Hastings scheme. Let θ represent the generic parameter, that is, any of $\boldsymbol{\beta}_1$, β_2, and $\boldsymbol{\gamma}$. Let $L_c(\theta|\cdot)$ denote the conditional likelihood as furnished in (8) as a function of θ given the data and all other parameters. Let $\pi(\theta)$ be the prior distribution on θ. To simulate observations from the full conditional distribution of θ, namely $\pi(\theta|\cdot)$, we followed these steps:

Step 1. Start with any reasonable initial value of θ, say θ_0. This is the current value for θ.

Step 2. Generate a new value of θ, say θ^*, from a candidate density $g(\theta)$.

Step 3. Replace θ_0 by θ^* with probability $\min\{1, \frac{\pi(\theta^*|\cdot)g(\theta_0)}{\pi(\theta_0|\cdot)g(\theta^*)}\}$. Retain the existing value of θ_0 otherwise.

Remark A.1. We chose the candidate density $g(\theta)$ to be identical to the prior density $\pi(\theta)$. Note that the full conditional density of θ, namely, $\pi(\theta|\cdot) \propto \pi(\theta)L_c(\theta|\cdot)$. Consequently, the acceptance probability as described in step 2 reduces to

$$\min\left\{1, \frac{L_c(\theta^*|\cdot)}{L_c(\theta_0|\cdot)}\right\}.$$

Once we update the value of the first set of parameters, say $\boldsymbol{\beta}_1$, we move on to the similar cycle for β_2, with the updated value of $\boldsymbol{\beta}_1$ substituted in the likelihood in (8). After updating β_2, we proceed to the Metropolis–Hastings step for $\boldsymbol{\gamma}_1$ with the updated values of $\boldsymbol{\beta}_1$ and β_2 and finally with all three updated values of $\boldsymbol{\beta}_1$, β_2, and $\boldsymbol{\gamma}_1$ to the iteration steps for γ_{0i}, $i = 1, \ldots, n$.

Remark A.2. For the equine example in Section 4.1 with the exposure (age) distribution modeled normally, there was an additional scale parameter, σ^2, for which we used an inverse gamma prior. In that case there is an additional Metropolis–Hastings step for σ^2 similar to that described earlier.

Generating Observations From the Conditional of γ_{0i}. For simulating observations from the full conditional of γ_{0i}, $i = 1, \ldots, n$, first note that this will be substantially computationally intensive, because n is typically large in many of the real examples (e.g., $n = 498$ for the equine data and $n = 63$ for the LA cancer study. We adopted one of the algorithms suggested by Neal (2000) for generating observations from the posterior of Dirichlet mixtures.

To illustrate the computational steps, we fix our attention to a particular stratum, say, stratum k. Also, let $\boldsymbol{\gamma}_{0(-k)} = (\gamma_{01}, \ldots, \gamma_{0k-1}, \gamma_{0k+1}, \ldots, \gamma_{0n})$, the vector of all but the kth stratum-effect parameter. We proceed in the following way:

Step 1. Start with some reasonable values of γ_{0i}, $i = 1, \ldots, n$. The current state of the Markov chain consists of these n values, namely $(\gamma_{01}, \ldots, \gamma_{0k}, \ldots, \gamma_{0n})$.

Step 2. Draw a candidate value γ^*_{0k} from the distribution of $\gamma_{0k}|\gamma_{0(-k)}$, namely from

$$(\alpha + n - 1)^{-1} \sum_{j \neq k}^{n} I(\gamma_{0k} = \gamma_{0j}) + \{\alpha/(\alpha + n - 1)\}\mathrm{N}(\zeta_0, \sigma_0^2).$$

There are several ways to generate observations from the foregoing mixture distribution, which essentially reflects that for the kth stratum effect you either get a new distinct value from the normal component of the mixture with probability $\alpha/(\alpha + n - 1)$ or otherwise it is drawn with equal probability from the current set of $(n - 1)$ entries of $\boldsymbol{\gamma}_{0(-k)}$, the vector corresponding to the other $(n - 1)$ stratum effect parameters. The following is one particular way to generate observations from this candidate density:

a. Let $p_j = 1/(\alpha + n - 1)$ for $j = 1, \ldots, n, j \neq k$ and $p_k = \alpha/(\alpha + n - 1)$.
b. Calculate the cumulative probabilities $c_l = \sum_{j=1}^{l} p_j$. Let $c_0 \equiv 0$.
c. Draw a random number u from U(0, 1) distribution. Then $c_{j-1} < u \leq c_j$ for some specific $j \in \{1, \ldots, n\}$. If $j = k$, then draw the candidate γ^*_{0k} from $\mathrm{N}(\zeta_0, \sigma_0^2)$. If $j \neq k$, then take $\gamma^*_{0k} = \gamma_{0j}$.

Step 3. Compute the acceptance probability $a(\gamma^*_{0k}, \gamma_{0k}) = \min\{1, L_c(\gamma^*_{0k}|\cdot)/L_c(\gamma_{0k}|\cdot)\}$, where $L_c(\gamma^*_{0k}|\cdot)$ and $L_c(\gamma_{0k}|\cdot)$ are the conditional likelihoods as given in (8) evaluated at γ^*_{0k} and γ_{0k}, with all other parameters held fixed at their current values.

Step 4. Set the new value of γ_{0k} to γ^*_{0k} with the foregoing probability.

Step 5. Repeat Steps 2–4 R times. Following the work of Neal (2000), we chose $R = 5$ in all our computations. Consider the last of these R updates of γ_{0k} as the current value of the kth stratum effect parameter. We repeat the foregoing steps for all n stratum-effect parameters, so for the equine data there are 498 such cycles to be performed, one for each of the γ_{0i}, $i = 1, \ldots, n$.

Remark A.3. For resampling from the full conditional distribution of α, we followed the algorithm suggested by Escobar and West (1995). At each step we counted the number of distinct γ_{0i}, say k, and conditional on the current values of α and k, we simulated a latent beta random variable, say, $\eta \sim \mathrm{B}(\alpha + 1, n)$. Let $G(a, b)$ denote the prior on α. Using the current value of k and η, we simulated α from the following mixture of gamma distribution:

$$\pi(\alpha|\eta, k) = pG\{a + k, b - \log(\eta)\} \\ + (1 - p)G\{a + k - 1, b - \log(\eta)\},$$

where $p = (a + k - 1)/[a + k - 1 + n\{b - \log(\eta)\}]$.

One complete step of the entire Markov chain consists of the foregoing componentwise Metropolis–Hastings steps for all the parameters in the model. We proceeded iteratively through the whole cycle for all of the parameters until convergence. We ran the chain typically from 7,000–10,000 iterations and calculated the diagnostic proposed by Gelman and Rubin (1992) as a measure of convergence. We furnished the posterior means as our estimates for the parameters and also provided the highest posterior density–credible region for all the parameters. Computer code for implementing the methodology is available on request.

[Received April 2003. Revised July 2004.]

REFERENCES

Antoniak, C. E. (1974), "Mixtures of Dirichlet Processes With Applications to Non-Parametric Problems," *The Annals of Statistics*, 2, 1152–1174.

Breslow, N. E. (1996), "Statistics in Epidemiology, the Case-Control Study," *Journal of the American Statistical Association*, 91, 14–28.

Breslow, N. E., and Day, N. E. (1980), *Statistical Methods in Cancer Research*, Vol. 1, Lyon, France: International Agency for Research on Cancer.

Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L., and Sabai, C. (1978), "Estimation of Multiple Relative Risk Functions in Matched Case-Control Studies," *American Journal of Epidemiology*, 108, 299–307.

Cohen, N. D. (1997), "Epidemiology of Colic," *Veterinary Clinic North America Equine Practice*, 13, 91–201.

Diggle, P. J., Morris, S. E., and Wakefield, J. C. (2000), "Point-Source Modeling Using Matched Case-Control Data," *Biostatistics*, 1, 89–105.

Escobar, M. D. (1994), "Estimating Normal Means With a Dirichlet Process Prior," *Journal of the American Statistical Association*, 89, 268–277.

Escobar, M. D., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Gelman, A., and Rubin, D. B. (1992), "Inference From Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, 457–472.

Ghosh, M., and Chen, M.-H. (2002), "Bayesian Inference for Matched Case-Control Studies," *Sankhyā*, Ser. B, 64, 107–127.

Hosmer, D. A., and Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.), New York: Wiley.

Kim, I., Cohen, N. D., and Carroll, R. J. (2002), "A Method for Graphical Representation of Effect Heterogeneity by a Matched Covariate in Matched Case-Control Studies Exemplified Using Data From a Study of Colic in Horses," *American Journal of Epidemiology*, 156, 463–470.

Lipsitz, S. R., Parzen, M., and Ewell, M. (1998), "Inference Using Conditional Logistic Regression With Missing Covariates," *Biometrics*, 54, 295–303.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data*, New York: Wiley.

MacEachern, S. N., and Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–238.

Müller, P., and Roeder, K. (1997), "A Bayesian Semiparametric Model for Case-Control Studies With Errors in Variables," *Biometrika*, 84, 523–537.

Neal, R. M. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

Paik, M. C., and Sacco, R. (2000), "Matched Case-Control Data Analyses With Missing Covariates," *Applied Statistics*, 49, 145–156.

Rathouz, P. J., Satten, G. A., and Carroll, R. J. (2002), "Semiparametric Inference in Matched Case-Control Studies With Missing Covariate Data," *Biometrika*, 89, 905–916.

Robert, C. P., and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer-Verlag.

Satten, G., and Carroll, R. J. (2000), "Conditional and Unconditional Categorical Regression Models With Missing Covariates," *Biometrics*, 56, 384–388.

Satten, G., Flanders, W. D., and Yang, Q. (2001), "Accounting for Unmeasured Population Substructure in Case-Control Studies of Genetic Association Using a Novel Latent-Class Model," *American Journal of Human Genetics*, 68, 466–477.

Satten, G., and Kupper, L. (1993a), "Conditional Regression Analysis of the Odds Ratio Between Two Binary Variables When One Is Not Measured With Certainty: A Method for Epidemiologic Studies," *Biometrics*, 49, 429–440.

——— (1993b), "Inferences About Exposure-Disease Associations Using Probability-of-Exposure Information," *Journal of the American Statistical Association*, 88, 200–208.

Seaman, S. R., and Richardson, S. (2001), "Bayesian Analysis of Case-Control Studies With Categorical Covariates," *Biometrika*, 88, 1073–1088.

Walsh, R. D. (1994), "Effects of Maternal Smoking on Adverse Pregnancy Outcomes: Examination of the Criteria of Causation," *Human Biology*, 66, 41–49.

Wang, C. Y., Wang, S., and Carroll, R. J. (1997), "Estimation in Choice-Based Sampling With Measurement Error and Bootstrap Analysis," *Journal of Econometrics*, 77, 65–86.

West, M., Müller, P., and Escobar, M. D. (1994), "Hierarchical Priors and Mixture Models, With Application in Regression and Density Estimation," in *Aspects of Uncertainty: A Tribute to D. V. Lindley*, eds. A. F. M. Smith and P. Freeman, New York: Wiley, pp. 363–386.

Zelen, M., and Parker, R. A. (1986), "Case-Control Studies and Bayesian Inference," *Statistics in Medicine*, 5, 261–269.