Sankhyā : The Indian Journal of Statistics
Special Issue on Statistics in Biology and Health Sciences
2007, Volume 69, Part 3, pp. 379-404

Modelling Association Among Bivariate Exposures In Matched Case-Control Studies

Samiran Sinha
Texas A&M University, College Station, USA
Bhramar Mukherjee
University of Michigan, Ann Arbor, USA
Malay Ghosh
University of Florida, Gainesville, USA

Abstract

The paper considers the problem of modelling association between two exposure variables in a matched case-control study, where both the exposures may be partially missing. The exposure variables could all be categorical or continuous or could be a mixed set of some categorical and some continuous variables. Association models for the missing exposure variables using the completely observed covariates and disease status are proposed for each of the three scenarios. The models account for varying stratum heterogeneity in different matched sets. Three real data examples accompany the proposed models. The examples as well as a small scale simulation study indicate that in presence of missingness and association, modelling the association between the exposures rather than ignoring it, often leads to better estimates of the relative risk parameters with smaller standard errors. Estimation of the model parameters is carried out in a Bayesian framework and the estimates are compared with classical conditional logistic regression estimates.

AMS (2000) *subject classification.* Primary 62F10, 62F15, 62H12.
Keywords and phrases. Association model, benign breast disease, colon cancer, conditional likelihood, endometrial cancer, matched designs, measurement errors, missing at random.

1 Introduction

The paper addresses inference problems based on matched case-control studies involving bivariate exposure variables, where both the variables could be partially missing. The main objective of a case-control study is to assess

the association between a certain outcome (typically a disease) and the exposure variables to ascertain potential risk factors for the disease. There are many instances, where the missing exposure variables may themselves be associated, and one of the key features of this paper is modelling this underlying association. In particular, we consider three association schemes: (i) two exposure variables with a bivariate binary distribution, (ii) a set of continuous exposure variables with an underlying correlation structure, and (iii) two correlated exposure variables, one binary and the other continuous. We propose a full-likelihood based approach by a parametric modelling of the missing exposure distribution incorporating their association. Our model accounts for possible stratum heterogeneity in the exposure distribution, which may be present in matched studies. Since the model and the resulting likelihood turn out to be of a complex form with a growing number of parameters, the parameters are estimated in a fully Bayesian framework by using the Markov chain Monte Carlo technique.

Outside the case-control context, there has been extensive amount of work on modelling association among multivariate exposures. Zhao and Prentice (1990) proposed a pseudo likelihood method for the estimation of parameters of a quadratic exponential family in terms of the marginal means and pairwise correlation. Lang et al. (1999) proposed association modelling in multivariate categorical response data. They formulated generalized log-linear models that simultaneously model the association structure as well as the marginal distributions of the responses and outlined how to obtain the maximum likelihood estimates of the parameters. Ekholm et al. (2000) proposed association models for multivariate binary data through latent variables and Markov type dependence structure. But so far, to the best of our knowledge, none of the models have been extended to the case-control domain.

The second feature of our model is to accommodate partial missingness in multiple exposure variables. Although the assumption is one of missing at random (MAR) (see Little and Rubin, 2002), our analysis is based not just on complete case data (i.e., data from subjects with complete information), but also on subjects carrying partial information. This is in contrast with the conditional logistic regression (CLR) model traditionally used for matched case-control problems, which uses only the complete case data.

Our method is neither similar to that of Lipsitz et al. (1998), and Rathouz et al. (2002), who proposed modelling the missingness process in matched case-control studies nor to that of Huberman and Langholz (1999), who used

missing-indicator method to handle incomplete data in matched studies. It is more akin to the work by Satten and Kupper (1993a, 1993b), Satten and Carroll (2000) and Paik and Sacco (2000), who proposed a likelihood based approach by parametrically modelling the distribution of the exposure variable.

Though there is a significant volume of literature on various aspects of case-control studies in the frequentist domain, Bayesian literature related to this area has been limited (see Mukherjee et al. (2005) for a complete review). Bayesian literature on case-control studies have mostly been on unmatched case-control studies (Zelen and Parker, 1986). Recently, Sinha et al. (2005) proposed a semiparametric Bayesian method of estimation in matched case-control scenario, when the exposure variable was partially missing. In Sinha et al. (2005), the association between a single missing exposure and a completely observed covariate was modelled through a linear regression but the model does not take into account the possibility of a dependence structure among the set of multivariate missing exposures themselves. Moreover, Sinha et al. (2005) consider exposure distributions belonging to the generalized exponential family. The current paper is not restricted to that class of distributions.

Three data-sets are analysed in the paper, each interesting in its own right. First, we consider the association structure for a bivariate binary exposure in the context of the well-known Los Angeles Endometrial Cancer Study (Breslow and Day, 1980), which involves natural missingness in one of the exposures. Second, we consider a data-set (Miller et al., 1983) relating dietary exposures like the total caloric and fibre intakes to incidence of colon cancer. In this example, we analyse missingness and measurement error simultaneously in a matched case-control study. Third, we consider the scenario when there is one binary, and one continuous exposure. A case-control data-set on fibrocystic breast disease (Pastides et al., 1985) is considered to illustrate the methodology. One of the significant exposures is partially missing in this data-set.

In all three examples, we find that by exploiting the association among the exposures, especially in the presence of missingness, we obtain better estimates with relatively smaller standard errors, where there is a real association among the exposures. A small scale simulation study supports this finding.

The rest of the article is organized as follows. In Section 2, we describe our model, notations and formulation of the likelihood in the presence of

missingness in a general matched case-control set-up. Section 3 is devoted to three different association models for multiple exposure variables as described above. Each model is accompanied by the corresponding data example illustrating the proposed methodology. Section 4 contains a small scale simulation study to indicate the effect of association modelling when association among exposures truly exists. Section 5 is the concluding section with a discussion of the findings and final remarks.

2 Model and Notations

We consider a $1 : M$ matched case-control design. For subject j in the matched set i, $j = 1, \ldots, M+1$; $i = 1, \ldots, s$, let D_{ij} be a binary indicator for the disease status, namely $D_{ij} = 1$ for case and $D_{ij} = 0$ for control. For each subject in the data-set, we observe a $p \times 1$ vector of exposure variables $\mathbf{X}_{ij}$ and a $q \times 1$ vector of covariates $\mathbf{Z}_{ij}$ which is completely observed. In all our examples, we considered $p = 2$. However, the proposed methodology in Section 3.2 works for arbitrary $p(\geq 2)$, and allows the possibility that different components of $\mathbf{X}_{ij}$ may be missing for different subjects. For each subject, missing values in the vector of exposure variables $\mathbf{X}_{ij}$ could occur in $K = 2^p$ ways. For example, if $\mathbf{X}_{ij}$ has two components (say $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$), possible patterns of missingness are (i) both $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ are missing (ii) $X_{ij}^{(1)}$ is missing and $X_{ij}^{(2)}$ is observed, (iii) $X_{ij}^{(2)}$ is missing and $X_{ij}^{(1)}$ is observed, and (iv) both $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ are observed. Let δ_{ij}^k, $k = 1, \ldots, K$, represent the indicator variables corresponding to each pattern of missingness. Therefore, for each subject, we observe a vector of indicator variables, namely $\Delta_{ij} = (\delta_{ij}^1, \delta_{ij}^2, \ldots \delta_{ij}^K)^T$, which takes value 1 in exactly one position and zero in all other positions. We assume that the missingness patterns are lexicographically ordered, i.e., $\delta_{ij}^1 = 1$ when the missingness pattern is $(0, 0, \ldots, 0)$ (denoting all exposures are missing), $\delta_{ij}^2 = 1$ if the missingness pattern is $(1, 0, \ldots, 0)$, $\delta_{ij}^{K-1} = 1$ if the pattern is $(0, 1, \ldots, 1)$, and $\delta_{ij}^K = 1$ if missingness pattern is $(1, 1, \ldots, 1)$. i.e., the exposure vector is completely observed.

Let $p(\mathbf{X}_{ij}|\mathbf{Z}_{ij}, D_{ij} = 0)$ be the density of the exposure variable in the control population with respect to a σ-finite dominating measure μ. Note that by modelling the distribution of $\mathbf{X}$ in terms of the completely observed covariate $\mathbf{Z}$, one is able to capture stratum heterogeneity measured through $\mathbf{Z}$, but there may still be some unexplained heterogeneity, which we would

model through a random intercept term. Let $p_0^k(\mathbf{X}_{ij})$ and $p_1^k(\mathbf{X}_{ij})$ denote respectively the joint densities of the observed components of $\mathbf{X}_{ij}$ corresponding to control and case populations conditional on $\mathbf{Z}_{ij}$, for $k = 1, 2, \ldots, K$. We omit the conditioning statement for simplicity in notation. When all the components of $\mathbf{X}_{ij}$ are missing we set $p_0^1(\mathbf{X}_{ij}) = 1$ and $p_1^1(\mathbf{X}_{ij}) = 1$. Also, we consider a prospective odds model for the disease given by $P(D_{ij} = 1|\mathbf{X}_{ij}, \mathbf{Z}_{ij})/P(D_{ij} = 0|\mathbf{X}_{ij}, \mathbf{Z}_{ij}) = \rho(\mathbf{X}_{ij}, \mathbf{Z}_{ij})$, where $\rho(\cdot)$ is a general non-negative function. The following two results (see Satten and Kupper, 1993b) hold without any specific distributional assumption.

LEMMA 1.

$$\frac{P(D_{ij} = 1|\mathbf{Z}_{ij})}{P(D_{ij} = 0|\mathbf{Z}_{ij})} = \int \rho(\mathbf{X}_{ij}, \mathbf{Z}_{ij}) p(\mathbf{X}_{ij}|\mathbf{Z}_{ij}, D_{ij} = 0) d\mu(\mathbf{X}_{ij}). \tag{2.1}$$

LEMMA 2.

$$p(\mathbf{X}_{ij}|\mathbf{Z}_{ij}, D_{ij} = 1) = \frac{p(\mathbf{X}_{ij}|\mathbf{Z}_{ij}, D_{ij} = 0)\rho(\mathbf{X}_{ij}, \mathbf{Z}_{ij})}{\int \rho(\mathbf{X}_{ij}, \mathbf{Z}_{ij}) p(\mathbf{X}_{ij}|\mathbf{Z}_{ij}, D_{ij} = 0) d\mu(\mathbf{X}_{ij})}. \tag{2.2}$$

Thus, by Lemma 2, one is able to derive the exposure distribution in the case population, from the prospective odds model and the exposure distribution in the control population. These lemmas will be utilized in deriving the likelihoods.

It is common to assume a prospective logistic model

$$P(D_{ij} = 1|\mathbf{X}_{ij}, \mathbf{Z}_{ij}) = H\{\beta_{0i} + \beta_1^T \mathbf{Z}_{ij} + \beta_2^T \mathbf{X}_{ij}\}, \tag{2.3}$$

where $H(u) = 1/\{1 + \exp(-u)\}$, β_{0i} is a stratum-specific intercept term, and β_1 and β_2 are the vectors of log-odds ratio parameters corresponding to the completely observed covariates $\mathbf{Z}_{ij}$ and the exposure variables $\mathbf{X}_{ij}$. In the case of fully observed data, the conditional likelihood (conditioned on $\sum_{j=1}^{M+1} D_{ij} = 1$) based on this prospective model, eliminates the nuisance parameters β_{0i} and involves only the log odds ratio parameters.

In the presence of missingness, the classical method of conditional logistic regression (CLR) ignores the partially observed records and drops completely observed records if there is no completely observed matching records. In case the data are missing completely at random (MCAR; for definition, see Little and Rubin, 2002), CLR produces consistent but less efficient estimates of the

parameters. On the other hand, if the data are missing at random (MAR), CLR may yield biased estimates (Breslow and Cain, 1988; Paik, 2004).

Sinha et al. (2005) proposed a Bayesian semiparametric method to account for missing data via modelling of the exposure distribution. The method proposed there does not incorporate the possible dependence structure among a set of multivariate exposure variables each with possible missingness. In Sinha et al. (2005), the information on any particular missing component of the exposure vector comes only through its association with the completely observed covariate Z, whereas with our current model, we also exploit the association of this particular missing component with the other partially observed components of the exposure vector X and thus gain in efficiency in case the association truly exists.

The basic structure of the joint likelihood in the presence of missingness is

$$p(D_{ij}, \mathbf{X}_{ij}, \Delta_{ij}|\mathbf{Z}_{ij}) = p(\mathbf{X}_{ij}|D_{ij}, \Delta_{ij}, \mathbf{Z}_{ij}) \times p(\Delta_{ij}|D_{ij}, \mathbf{Z}_{ij}) \times p(D_{ij}|\mathbf{Z}_{ij}). \tag{2.4}$$

One should note that there is an underlying asymmetry in the model. We treat the partially missing exposure $\mathbf{X}$ as stochastic, and assume a parametric distribution for it. On the other hand, we consider the completely observed covariate Z as non-stochastic. Theoretically, one can always model Z with some parametric distribution. However, the Z's are usually multivariate with a mixture of categorical and continuous covariates, so that specifying a true model for its distribution is not an easy task; needless to say, if the specified model is incorrect, we risk the possibility of biased and inconsistent estimation.

Following Satten and Carroll (2000), we will assume that (i) the $\mathbf{X}$'s are missing at random, i.e., $p(\mathbf{X}_{ij}|D_{ij}, \Delta_{ij}, \mathbf{Z}_{ij}) = p(\mathbf{X}_{ij}|D_{ij}, \mathbf{Z}_{ij})$, and (ii) $p(\Delta_{ij}|D_{ij}, \mathbf{Z}_{ij})$ does not depend on the log-odds ratio parameters (i.e., β_1 and β_2).

By Lemmas 1 and 2, and assuming without loss of generality that the first subject in each stratum is a case, and the rest are controls, one begins with the joint conditional likelihood

$$L_c(\cdot) = \prod_{i=1}^{s} P\Bigg(D_{i1} = 1, D_{i2} = 0, \ldots, D_{iM+1} = 0, \{\mathbf{X}_{ij}, \Delta_{ij}\}_{j=1}^{M+1} \Bigg| \{\mathbf{Z}_{ij}\}_{j=1}^{M+1}, \sum_{r=1}^{M+1} D_{ir} = 1 \Bigg),$$

which can be written as

$$
\begin{aligned}
L_c(\cdot) \;\propto\; & \prod_{i=1}^{s}\Bigg\{P\Bigg(D_{i1}=1, D_{i2}=0,\ldots,D_{iM+1}=0, \Bigg|\{\mathbf{Z}_{ij}\}_{j=1}^{M+1}, \sum_{r=1}^{M+1} D_{ir}=1\Bigg) \\
& \times p(\mathbf{X}_{i1}|\mathbf{Z}_{i1}, D_{i1}=1, \Delta_{i1}) \prod_{j=2}^{M+1} p(\mathbf{X}_{ij}|\mathbf{Z}_{ij}, D_{ij}=0, \Delta_{ij})\Bigg\} \\
\propto\; & \prod_{i=1}^{s} \frac{P(D_{i1}=1|\mathbf{Z}_{i1})/P(D_{i1}=0|\mathbf{Z}_{i1})}{\sum_{j=1}^{M+1} P(D_{ij}=1|\mathbf{Z}_{ij})/P(D_{ij}=0|\mathbf{Z}_{ij})} \\
& \times \prod_{i=1}^{s}\Bigg\{\Bigg(\sum_{k=1}^{K}\delta_{i1}^{k}p_1^k(\mathbf{X}_{i1})\Bigg) \times \prod_{j=2}^{M+1}\Bigg(\sum_{k=1}^{K}\delta_{ij}^{k}p_0^k(\mathbf{X}_{ij})\Bigg)\Bigg\}. \qquad (2.5)
\end{aligned}
$$

Note that, instead of using a retrospective likelihood, we work with a joint conditional likelihood of disease variable, exposure variable and missing value indicator given the completely observed covariates, the stratum effects and the number of cases in each matched set, which is held fixed by the matched design.

We may note that the likelihood given in (2.4) does not require the logistic assumption as given in (2.3), though we will assume the same in our actual applications in the subsequent sections. The parameters involved in the above likelihood are estimated in a fully Bayesian framework. As we will find, there will be no closed form available for the posterior and any exact Bayesian inference is analytically intractable. Hence, we estimate the model parameters through Markov chain Monte Carlo numerical integration. We use standard Metropolis-Hastings algorithm to generate random number from the conditional distributions (Robert and Casella, 1999).

In the following sections, we present the specific forms of the likelihoods for the three association scenarios we consider.

3 Three Different Association Models

3.1. Bivariate binary exposure variables. We first consider the situation with bivariate binary exposure variables $\mathbf{X}_{ij} = (X_{ij}^{(1)}, X_{ij}^{(2)})^T$. Assume that the retrospective density of $\mathbf{X}_{ij}$ in the control population to be

$$
p(X_{ij}^{(1)}, X_{ij}^{(2)}|\mathbf{Z}_{ij}, D_{ij}=0) = \frac{1}{C_{ij}^{(0)}} \exp(\theta_{ij}^{(1)}X_{ij}^{(1)} + \theta_{ij}^{(2)}X_{ij}^{(2)} + \lambda_{ij}X_{ij}^{(1)}X_{ij}^{(2)}), \qquad (3.1)
$$

where $C_{ij}^0 = 1 + \exp(\theta_{ij}^{(1)}) + \exp(\theta_{ij}^{(2)}) + \exp(\theta_{ij}^{(1)} + \theta_{ij}^{(2)} + \lambda_{ij})$. We model $\theta_{ij}^{(1)}$, $\theta_{ij}^{(2)}$, and λ_{ij} as $\theta_{ij}^{(1)} = \gamma_{01i} + \gamma_{11}\mathbf{Z}_{ij}$, $\theta_{ij}^{(2)} = \gamma_{02i} + \gamma_{12}\mathbf{Z}_{ij}$, and $\lambda_{ij} = \gamma_{03i} + \gamma_{13}\mathbf{Z}_{ij}$.

REMARK 1. Note that λ_{ij} is the log-odds ratio between the two random variables $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$, i.e.,

$$\lambda_{ij} = \log \frac{p(X_{ij}^{(1)}=1, X_{ij}^{(2)}=1|\mathbf{Z}_{ij}, D_{ij}=0)p(X_{ij}^{(1)}=0, X_{ij}^{(2)}=0|\mathbf{Z}_{ij}, D_{ij}=0)}{p(X_{ij}^{(1)}=0, X_{ij}^{(2)}=1|\mathbf{Z}_{ij}, D_{ij}=0)p(X_{ij}^{(1)}=1, X_{ij}^{(2)}=0|\mathbf{Z}_{ij}, D_{ij}=0)}.$$

Here $\lambda_{ij} = 0$ implies that the exposure variables are independent in the control population, and if $\lambda_{ij} \neq 0$ then the exposures are associated among the controls. Larger the departure from 0 stronger is the association.

Also, $\theta_{ij}^{(1)}$ could be interpreted as the logit of the probability of success of $X_{ij}^{(1)}$ conditioned on $X_{ij}^{(2)} = 0$, i.e., $p(X_{ij}^{(1)}|\mathbf{Z}_{ij}, D_{ij} = 0, X_{ij}^{(2)} = 0) = \exp(\theta_{ij}^{(1)} X_{ij}^{(1)})/\{1 + \exp(\theta_{ij}^{(1)})\}$. Similarly, $\theta_{ij}^{(2)}$ could be interpreted as the logit of the probability of success of $X_{ij}^{(2)}$ given that $X_{ij}^{(1)} = 0$.

Let $\beta_2 = (\beta_{21}, \beta_{22})^T$. By Lemma 1, (2.4) and (3.1), one obtains

$$\frac{P(D_{ij}=1|\mathbf{Z}_{ij})}{P(D_{ij}=0|\mathbf{Z}_{ij})} = \exp\{\beta_{0i} + \beta_1^T \mathbf{Z}_{ij} + \log(C_{ij}^{(1)}/C_{ij}^{(0)})\}, \tag{3.2}$$

where $C_{ij}^{(1)} = 1 + \exp(\theta_{ij}^{(1)} + \beta_{21}) + \exp(\theta_{ij}^{(2)} + \beta_{22}) + \exp(\theta_{ij}^{(1)} + \theta_{ij}^{(2)} + \lambda_{ij} + \beta_{21} + \beta_{22})$.

In this situation, there are four possible missing patterns leading to $\Delta_{ij} = (\delta_{ij}^1, \delta_{ij}^2, \delta_{ij}^3, \delta_{ij}^4)^T$. Using (2.4) and (3.2), we can write the likelihood as

$$L_c \propto \prod_{i=1}^{s} \left\{ \frac{\exp\{\beta_1^T \mathbf{Z}_{i1} + log(C_{i1}^{(1)}/C_{i1}^{(0)})\}}{\sum_{j=1}^{M+1} \exp\{\beta_1^T \mathbf{Z}_{ij} + log(C_{ij}^{(1)}/C_{ij}^{(0)})\}} \times \sum_{k=1}^{4} \delta_{i1}^k p_1^k(\mathbf{X}_{i1}) \times \prod_{j=2}^{M+1} (\sum_{k=1}^{4} \delta_{ij}^k p_0^k(\mathbf{X}_{ij})) \right\},$$

where $p_1^k(\mathbf{X}_{ij})$ and $p_0^k(\mathbf{X}_{ij})$ are the control and the case densities of the exposures respectively. The above likelihood involves β_1, β_2, γ_{11}, γ_{12}, γ_{13}, and γ_{0pi}, for $p = 1, 2, 3$ and $i = 1, \ldots, s$. We use independent normal

prior Normal$(0, 10)$ on each set of parameters, and estimate them through a Markov chain Monte Carlo integration scheme.

Example 1. Endometrial Cancer Data. We consider matched case-control data from the well-known Los Angeles Endometrial Cancer Study (Breslow and Day, 1980). The cases were matched with four controls on the basis of age, the community the case lived in, marital status, and the time the case had entered in the community. The data contains measurements on several binary covariates: presence of gall bladder disease, obesity, use of estrogen, to name a few. Several studies have indicated that exogenous estrogen treatment increases the risk of endometrial cancer (see Schottenfeld and Fraumeni, 1996). Among other factors, obesity may increase the risk of endometrial cancer by altering estrogen metabolism. The presence of gall bladder disease may increase the risk of endometrial cancer and that could potentially be due to an association of gall bladder disease with obesity or estrogen replacement therapy. With this underlying medical theory in mind, we reanalysed the LA cancer data-set with obesity and gall bladder disease as two associated binary exposures. Obesity has 16% missing values in this data-set. Estrogen plays the role of a completely observed covariate Z through which parameters related to the bivariate binary exposure distribution are modelled.

Bivariate binary distribution is assumed for the two exposure variables, obesity and gall bladder disease in the control population. The related parameters $\theta_{ij}^{(1)}$, $\theta_{ij}^{(2)}$ and λ_{ij}'s are modelled in terms of Z (estrogen use) as: $\theta_{ij}^{(1)} = \gamma_{01i} + \gamma_{11} Z_{ij}$, $\theta_{ij}^{(2)} = \gamma_{02i} + \gamma_{12} Z_{ij}$ and $\lambda_{ij} = \gamma_{03i} + \gamma_{13} Z_{ij}$. We used independent Normal(0,10) priors for all these regression parameters as well as for β_{21} and β_{22}. For each of our examples, we performed three analyses. The first is our proposed method of Bayesian analysis accounting for the association among the exposure variables (denoted by AM). The second is a Bayesian method following Sinha et al. (2005), where the missing exposure variables are assumed to be independent but each component of the exposure variable is related with the completely observed covariate Z through a linear regression (this model is denoted by IM) and the third one is the usual conditional logistic regression (CLR).

We carried out the three analyses on the observed data. We reran all the three methods when 30% observations on the exposure variable (the presence of gall bladder disease) were deleted completely at random. We then compared the performance of the three methods in the presence and absence of missingness.

Table 1. Results of the Endometrial Cancer Data example

Method		Estrogen-Use	Gall-Bladder	Obesity	CPO
Results with no missing data					
AM	Mean	2.18	1.11	0.67	–336.82
	s.e.	0.42	0.40	0.39	
	Lower HPD	1.48	0.35	–0.24	
	Upper HPD	3.15	1.94	1.40	
IM	Mean	2.23	1.11	0.65	–335.57
	s.e.	0.45	0.41	0.41	
	Lower HPD	1.42	0.35	–0.20	
	Upper HPD	3.20	1.95	1.43	
CLR	Mean	1.95	1.26	0.42	
	s.e.	0.50	0.44	0.40	
	Lower CL	0.98	0.41	–0.36	
	Upper CL	2.93	2.11	1.19	
Results with 30% Artificial Missingness on Gall-Bladder					
AM	Mean	2.37	1.31	0.75	–379.72
	s.e.	0.45	0.50	0.40	
	Lower HPD	1.56	0.25	–0.28	
	Upper HPD	3.32	2.42	1.50	
IM	Mean	2.52	1.70	0.64	–378.50
	s.e.	0.49	0.56	0.42	
	Lower HPD	0.66	0.68	–0.22	
	Upper HPD	3.45	2.86	1.40	
CLR	Mean	1.69	0.87	0.07	
	s.e.	0.64	0.58	0.55	
	Lower CL	0.45	–0.27	–1.00	
	Upper CL	2.94	2.01	1.15	

The summary of results is given in Table 1. Here, "Mean" is the posterior mean, "s.e." is the posterior standard deviation, "Lower HPD" and "Upper HPD" are the lower and upper end of the HPD region respectively,"Lower CL" and "Upper CL" are the lower and the upper ends of the confidence limit respectively. There is 16% natural missingness in obesity. For the observed data, we notice that use of estrogen and gall bladder disease are both significant risk factors for endometrial cancer. Although there is very little numerical difference between the estimates obtained from the AM and IM models, the AM model yields estimates with slightly smaller standard error. In contrast, for CLR we observe significantly higher standard errors both in complete and missing data situations. With artificial missingness in gall bladder disease, we notice a significant difference in the estimates, and in the standard errors between AM and IM. AM produces estimates with

smaller standard error than IM, and the AM estimates are relatively closer to their original observed data counterparts. As expected, CLR produces worse estimates in the presence of missing exposure variable.

Table 2 gives the estimates of the other regression parameters present in the AM and IM model. Here, "Mean" is the posterior mean, "S.E." is the posterior standard deviation. Missing data means the data with 30% artificial missingness on the gall-bladder disease. The estimate of the parameter γ_{13} suggests that the association between the two exposures is moderately weak for this data-set.

TABLE 2. SECONDARY MODEL PARAMETER ESTIMATES FOR ASSOCIATION AND INDEPENDENCE MODELLING OF THE ENDOMETRIAL CANCER DATA EXAMPLE

Parameter		Association modelling		Independence modelling	
		Full Data	Missing Data	Full Data	Missing Data
γ_{11}	Mean	–1.40	–1.81	–1.17	–2.07
	S. E.	0.45	0.53	0.35	0.49
γ_{12}	Mean	0.19	0.26	0.22	0.25
	S.E.	0.30	0.31	0.30	0.32
γ_{13}	Mean	0.44	0.33		
	S.E.	0.53	0.63		

3.2. Multiple continuous exposures. In this section we consider a vector of multiple continuous exposure variables $\mathbf{X}_{ij}$. It is assumed that $\mathbf{X}_{ij}$ has a multivariate normal distribution in the control population, i.e.,

$$[\mathbf{X}_{ij}|\mathbf{Z}_{ij}, D_{ij} = 0] \sim \text{Normal}(\theta_{ij}, \Sigma), \tag{3.3}$$

where $\theta_{ij} = \gamma_{0i} + \gamma_1 \mathbf{Z}_{ij}$. Now using (3.3) in Lemma 1, one obtains

$$\begin{aligned} \frac{P(D_{ij}=1|\mathbf{Z}_{ij})}{P(D_{ij}=0|\mathbf{Z}_{ij})} &= \int \exp\{\beta_{0i} + \beta_1^T \mathbf{Z}_{ij} + \beta_2^T \mathbf{X}_{ij}\} \\ &\quad \times |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{X}_{ij} - \theta_{ij})^T \Sigma^{-1} (\mathbf{X}_{ij} - \theta_{ij})\} d\mathbf{X}_{ij} \\ &= \exp\{\beta_{0i} + \beta_1^T \mathbf{Z}_{ij} + \beta_2^T \theta_{ij} + \frac{1}{2}\beta_2^T \Sigma \beta_2\}. \end{aligned} \tag{3.4}$$

The joint conditional likelihood is obtained as

$$L_c \propto \prod_{i=1}^{s} \left\{ \frac{\exp\{\beta_1^T \mathbf{Z}_{i1} + \beta_2^T \theta_{i1}\}}{\sum_{j=1}^{M+1} \exp\{\beta_1^T \mathbf{Z}_{ij} + \beta_2^T \theta_{ij}\}} \times \sum_{k=1}^{K} \delta_{i1}^k p_1^k(\mathbf{X}_{i1}) \times \prod_{j=2}^{M+1} \left(\sum_{k=1}^{K} \delta_{ij}^k p_0^k(\mathbf{X}_{ij}) \right) \right\}. \tag{3.5}$$

Often we face the problem of handling measurement error in measuring continuous exposure variables. Examples can be found in Carroll et al. (1993), Roeder et al. (1996), Gustafson et al. (2000, 2002) among others. In nutrition studies, for example, measurement error occurs naturally in recording data on a person's dietary intake. Since our second example is one in which we need adjustment for measurement error in the values of multiple continuous exposures, we briefly discuss an adjustment method we used to deal with measurement error and missingness simultaneously.

Assume that we observe an error-prone variable $\mathbf{T}$ instead of the true values of the exposure $\mathbf{X}$. We assume an additive error structure $\mathbf{T}_{ij} = \mathbf{X}_{ij} + \mathbf{U}_{ij}$, $j = 1, \ldots, M+1$, $i = 1, \ldots, s$, and also assume that conditional on $\mathbf{X}_{ij}$, the disease variable D_{ij} is independent of $\mathbf{T}_{ij}$. Further, we assume the non-differential measurement error model (as described in Carroll and Stefanski, 1994),

$$[\mathbf{U}_{ij}|\mathbf{X}_{ij}] \sim \text{Normal}(\mathbf{C}, \Sigma_u). \tag{3.6}$$

As stated earlier, we assume a multivariate normal distribution of the exposure variable in the control population. Note that the first factor of (3.5) does not involve $\mathbf{X}_{ij}$. Hence,we should only replace the other two factors in (3.5) by the distributions of $\mathbf{T}_{ij}$, $j = 1, \ldots, M+1$.

Using additive error structure, (3.3), (3.4) and (3.6), one can obtain the distributions of $\mathbf{T}$ in the control and case populations as

$$[\mathbf{T}_{ij}|\mathbf{Z}_{ij}, D_{ij} = 0] \sim \text{Normal}(\mathbf{C} + \theta_{ij}, \Sigma + \Sigma_u); \tag{3.7}$$

$$[\mathbf{T}_{ij}|\mathbf{Z}_{ij}, D_{ij} = 1] \sim \text{Normal}(\mathbf{C} + \theta_{ij} + \Sigma\beta_2, \Sigma + \Sigma_u). \tag{3.8}$$

Let $p_1^k(\mathbf{T}_{ij})$ and $p_0^k(\mathbf{T}_{ij})$ be the densities of observed components of $\mathbf{T}_{ij}$ in case and control population respectively corresponding to missingness patterns $k = 1, \ldots, K$. So the modified likelihood in the presence of measurement error and missingness is

$$\begin{aligned} L_c \;=\; & \prod_{i=1}^{s} \left\{ P\left(D_{i1} = 1, D_{i2} = 0, \ldots, D_{iM+1} = 0 \,\middle|\, \{\mathbf{Z}\}_{j=1}^{M+1}, \sum_{j=1}^{M+1} D_{ij} = 1 \right) \right. \\ & \left. \times p(\mathbf{T}_{i1}|\mathbf{Z}_{i1}, \Delta_{i1}, D_{i1} = 1) \times \prod_{j=2}^{M+1} p(\mathbf{T}_{ij}|\mathbf{Z}_{ij}, \Delta_{ij} D_{ij} = 0) \right\}, \end{aligned}$$

which can be written as

$$L_c \propto \prod_{i=1}^{s} \left\{ \frac{\exp(\beta_1^T \mathbf{Z}_{i1} + \beta_2^T \theta_{i1})}{\sum_{r=1}^{M+1} \exp(\beta_1^T \mathbf{Z}_{ir} + \beta_2^T \theta_{ir})} \times \sum_{k=1}^{K} \delta_{i1}^k p_1^k(\mathbf{T}_{i1}) \times \prod_{j=2}^{M+1} \left(\sum_{k=1}^{K} \delta_{ij}^k p_0^k(\mathbf{T}_{ij}) \right) \right\}. \tag{3.9}$$

The likelihood in (3.9) involves β_1, β_2, γ_{0i}, γ_1 along with $\mathbf{C}$, and Σ_u of the error distribution. We replace $\mathbf{C}$ and Σ_u by their estimates obtained from a validation data-set. We use independent normal priors for the components of β_1, β_2, γ_{0i} and γ_1. We write $\Sigma = ((\sigma_{ij})) = \mathbf{VRV}$, where $\mathbf{R}$ is the correlation matrix, and $\mathbf{V} = \text{Diag}(\sigma_{11}^{1/2}, \ldots, \sigma_{pp}^{1/2})$.

For our example, we consider only two exposures and assume a uniform (–1, 1) prior on their correlation coefficient and inverse gamma priors on the variance components of the exposure variables. When we consider more than two exposures, we will have to assume a prior ensuring a valid multivariate covariance matrix (Barnard et al., 2000). Finally, we run a Metropolis-Hastings algorithm to generate random numbers from the posterior distribution of the parameters.

REMARK 2. Often it is more appropriate to jointly model the validation data (which allows us to estimate the error distribution) and the primary data to reduce the bias of estimation. Since our main goal here is illustrating the proposed method of handling association, and not to develop any novel method of handling measurement error, we use the naive method to get information about $\mathbf{C}$ and Σ_u from the validation data as in Armstrong et al. (1989).

EXAMPLE 2. COLON CANCER STUDY. Here we consider a matched case-control data from a colon cancer study in Canada (see Miller et al., 1983), where 171 male cases of colon cancer were individually matched by age and neighbourhood of residence to 171 controls. Each subject's data is based on a quantitative diet history questionnaire referring to two-month period antedating the diagnosis of colon cancer in the cases, and a corresponding time period in the controls. Height and weight are measured for each study individual apart from their measurement on intake of calories, protein, total fiber, and R carotene. The goal of the study is to see the effect of the two

continuous exposure variables, namely total calories ($X^{(1)}$) and fiber intake ($X^{(2)}$) on the risk of colon cancer. An interesting feature of this data-set is that reliability and validity studies reveal that the measurement on the exposure variables are subject to a correlated measurement error structure.

Armstrong et al. (1989) dealt with the issue of measurement error in this data-set. We propose a Bayesian alternative in order to correct for the measurement error and then model the association of the adjusted exposures in terms of completely observed covariates such as height and weight of the subject. We also study the effect of missingness in this situation by introducing artificial missingness in the exposures.

There was an initial study, which investigated the validity of the data by comparing dietary histories reported by 16 healthy volunteers with detailed weighed food records kept by the volunteers' spouse. The spousal records were considered as a gold standard. The sample mean and the sample dispersion of the difference between these two records served as a direct estimate of the parameters of the distribution of the measurement error. Therefore the estimate of the mean of $\mathbf{U}_{ij}$ namely, $\mathbf{C}$ and its dispersion matrix Σ_u, are obtained from this validation data. We consider height and weight as two covariates.

Sample means of recorded and reported intakes for total calories and fiber from the validation data were $\bar{\mathbf{T}} = (32.0, 31.9)$ and $\bar{\mathbf{X}} = (25, 20.4)$; their mean difference serves as an estimate of $\mathbf{C}$, namely, $\hat{\mathbf{C}} = (6.9, 11.5)$. The moment estimate of Σ_u from the validation data is

$$\hat{\Sigma}_u = \begin{pmatrix} 56.6 & 35.1 \\ 35.1 & 70.1 \end{pmatrix},$$

and consequently, Σ_u is replaced by $\hat{\Sigma}_u$ in the likelihood in (3.9).

For the association model, we used independent Normal(0,10) priors for the components of β_1, β_2, γ_{0i} and γ_1. We used inverse gamma priors with scale parameter 100 and shape parameters 3 and 2.9 respectively on the variance components of $\mathbf{X}$. For the independence modelling, we set the correlation coefficient between the two exposure variables to zero (i.e., we assume both Σ and Σ_u to be diagonal) in the above likelihood and carry out our analysis. For the conditional logistic regression analysis, we first obtain the estimates by CLR method and then make necessary correction for the measurement error in the exposure variables as suggested by Armstrong et al. (1989).

TABLE 3. RESULTS OF THE COLON CANCER DATA EXAMPLE

Method		Height	Weight	Calory	Dietary	$\sigma^2_{X_1}$	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$	CPO
Results with no missing data									
AM	Mean	0.42	0.58	0.064	–0.049	39.53	59.06	0.37	
	s.e.	0.82	0.85	0.031	0.026	7.78	9.56	0.09	–250.07
	Lower HPD	–1.15	–1.15	0.02	–0.10	24.10	40.50	0.19	
	Upper HPD	2.05	2.21	0.15	–0.01	54.95	77.89	0.56	
IM	Mean	0.34	0.70	0.065	–0.01	38.93	61.31		
	s.e.	0.84	0.79	0.021	0.021	6.82	9.99		–272.92
	Lower HPD	–1.12	–1.00	0.02	–0.04	24.90	44.20		
	Upper HPD	2.25	2.01	0.10	0.03	53.15	82.90		
CLR	Mean	0.39	0.57	0.098	–0.036				
	s.e.	0.89	0.86	0.034	0.021				
	Lower CL	–1.35	–1.13	0.031	–0.077				
	Upper CL	2.13	2.25	0.165	0.005				
Results with 20% missingness on Calory and Dietary Fiber									
AM	Mean	0.50	0.63	0.089	–0.036	23.77	43.22	0.09	
	s.e.	0.83	0.87	0.032	0.033	5.40	8.37	0.16	–300.19
	Lower HPD	–1.16	–1.14	0.04	–0.10	15.00	25.88	–0.25	
	Upper HPD	2.09	2.25	0.16	0.02	35.50	58.80	0.40	
IM	Mean	0.43	0.55	0.086	–0.046	23.95	41.62		
	s.e.	0.86	0.80	0.047	0.035	5.29	7.81		–329.87
	Lower HPD	–1.12	–1.02	0.01	–0.08	14.10	29.21		
	Upper HPD	2.25	1.95	0.17	0.01	36.07	61.56		
CLR	Mean	2.16	–1.95	0.093	–0.119				
	s.e.	1.55	1.46	0.058	0.043				
	Lower CL	–0.88	–4.81	–0.021	–0.203				
	Upper CL	5.20	0.91	0.207	–0.035				

The results are presented in Table 3. Here, "Mean" is the posterior mean, "s.e." is the posterior standard deviation, "Lower HPD" and "Upper HPD" are the lower and the upper end of the HPD region respectively,"Lower CL" and "Upper CL" are the lower and the upper end of the confidence limit respectively. The CLR estimate of the exposures are obtained after the correction proposed by Armstrong et al. (1989). Once again, we notice that with the introduction of artificial missingness, the AM model has an edge over the IM model.

3.3. One binary exposure and the other belonging to an exponential family. In this section, we consider the modelling of a mixed set of continuous and categorical exposures. Though we write our model with the categorical

exposure to be binary, the methods may easily be extended to any categorical exposure.

Suppose that $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ are two exposure variables for the j^{th} subject in the i^{th} stratum. We assume that $X_{ij}^{(1)}$ is binary and the conditional density of $X_{ij}^{(2)}$ given $X_{ij}^{(1)}$ belongs to a general exponential family model for all i and j. Let

$$p(X_{ij}^{(1)} = 1|\mathbf{Z}_{ij}, D_{ij} = 0) = \pi_{ij}, \tag{3.10}$$

where we model $\text{logit}(\pi_{ij}) = \gamma_{01i} + \gamma_1^T \mathbf{Z}_{ij}$. The conditional distribution of $X_{ij}^{(2)}$ given $X_{ij}^{(1)}$ is assumed to be

$$p(X_{ij}^{(2)}|\mathbf{Z}_{ij}, D_{ij} = 0, X_{ij}^{(1)} = 0) = \exp\left\{\frac{\theta_{0ij}X_{ij}^{(2)} - b(\theta_{0ij})}{\phi_0} + c(X_{ij}^{(2)}, \phi_0)\right\} \tag{3.11}$$

$$p(X_{ij}^{(2)}|\mathbf{Z}_{ij}, D_{ij} = 0, X_{ij}^{(1)} = 1) = \exp\left\{\frac{\theta_{1ij}X_{ij}^{(2)} - b(\theta_{1ij})}{\phi_1} + c\left(X_{ij}^{(2)}, \phi_1\right)\right\} \tag{3.12}$$

where $\theta_{0ij} = \gamma_{02i} + \gamma_2^T \mathbf{Z}_{ij}$ and $\theta_{1ij} = \gamma_{03i} + \gamma_3^T \mathbf{Z}_{ij}$. This implies that the marginal distribution of $X_{ij}^{(2)}$ is a two component mixture of distributions that belongs to the exponential family for both the case and the control populations.

Remark 3. In the special case when $X_{ij}^{(2)}$ is a binary exposure variable with the parametrization,

$$p(X_{ij}^{(2)}|\mathbf{Z}_{ij}, D_{ij} = 0, X_{ij}^{(1)} = 0) = \exp\{\theta_{0ij}X_{ij}^{(2)} - \log(1 + \exp(\theta_{0ij}))\} \tag{3.13}$$

$$p(X_{ij}^{(2)}|\mathbf{Z}_{ij}, D_{ij} = 0, X_{ij}^{(1)} = 1) = \exp\{\theta_{1ij}X_{ij}^{(2)} - \log(1 + \exp(\theta_{1ij}))\} \tag{3.14}$$

and $p(X_{ij}^{(1)}|\mathbf{Z}_{ij}, D_{ij} = 0) = \exp\{\xi_{ij}X_{ij}^{(1)} - \log(1 + \exp(\xi_{ij}))\}$, the joint distribution of $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ is the bivariate binary distribution (3.1) of Section 3.1 with $\theta_{ij}^{(1)} = \xi_{ij} + \log\{(1 + \exp(\theta_{0ij}))/(1 + \exp(\theta_{1ij}))\}$, $\theta_{ij}^{(2)} = \theta_{0ij}$ and $\lambda_{ij} = \theta_{1ij} - \theta_{0ij}$.

The joint conditional likelihood is obtained as

$$L_c \propto \prod_{i=1}^{s}\left\{\frac{\exp(\beta_1^T \mathbf{Z}_{i1})g_{i1}}{\sum_{j=1}^{M+1}\exp(\beta_1^T \mathbf{Z}_{ij})g_{ij}} \times \sum_{k=1}^{4}\delta_{i1}^k p_1^k(\mathbf{X}_{i1}) \times \prod_{j=2}^{M=1}\left(\sum_{k=1}^{4}\delta_{ij}^k p_0^k(\mathbf{X}_{ij})\right)\right\}, \tag{3.15}$$

where

$$g_{ij} = (1-\pi_{ij})\exp\left\{\frac{b(\theta_{0ij}+\phi_0\beta_{22})-b(\theta_{0ij})}{\phi_0}\right\} + \pi_{ij}\exp\left\{\frac{b(\theta_{1ij}+\phi_1\beta_{22})-b(\theta_{1ij})}{\phi_1}+\beta_{21}\right\}. \quad (3.16)$$

Independent Normal(0,10) priors were assigned to the different components of β_1, β_2, γ_1, γ_2,γ_3, and γ_{0i}. Note that the extension of this method to the situation when conditional distribution of $X_{ij}^{(2)}$ given $X_{ij}^{(1)}$ belongs to a multi-parameter generalized exponential family, is straightforward and does not need any extra calculation. However, extension of this model is restricted to a general categorical exposure $X_{ij}^{(1)}$ with a few categories only, and it does not apply to count data.

EXAMPLE 3. BENIGN BREAST DISEASE DATA. This data is part of a large case-control study done in 1979 at two hospitals in New Haven, Connecticut (see Pastides et al., 1985) to assess the risk factors for benign fibrocystic breast diseases. The original data consist of 255 women (cases) aged 20-74 years with biopsy confirmed fibrocystic breast lesions and 790 controls at two hospitals in New Haven, Connecticut during 1979 (Pastides et al., 1985). The purpose of the study was to ascertain whether known risk factors for malignant breast cancer are also established risk factors for this type of benign breast diseases. There are many variables recorded in this data-set including demographic characteristics of the patient, medical history information and history of breast cancer in the family. The fraction of the data used in our analysis consists of 50 strata each of which contains 1 case and 3 controls. Controls were matched with a case on the basis of their age at the time of interview. The data-set was discussed in Section 7.8 of Hosmer and Lemeshow (2000). Cases were found to have a significantly higher level of education, a recent history of medical check-ups and a higher age at first pregnancy. The latter two exposures have natural missingness in the data-set. For the purpose of our analysis, we consider education as the completely observed covariate Z. History of medical checkups and age at first pregnancy are considered as the two exposure variables, say $X^{(1)}$ and $X^{(2)}$, both containing natural missingness. Here Z and $X^{(1)}$ are defined specifically as follows:

$$Z_{ij} = \begin{cases} 1 & \text{if the } ij^{th} \text{ subject has a junior college degree or above,} \\ 0 & \text{otherwise;} \end{cases}$$

$$X_{ij}^{(1)} = \begin{cases} 1 & \text{if the } ij^{th} \text{ subject had regular medical checkups,} \\ 0 & \text{otherwise.} \end{cases}$$

We model $\log\{\pi_{ij}/(1-\pi_{ij})\} = \gamma_{01i} + \gamma_1 Z_{ij}$, and the canonical parameters involved in the distributions of $X^{(2)}$ are modelled as $\theta_{0ij} = \gamma_{02i} + \gamma_2 Z_{ij}$ and $\theta_{ij} = \gamma_{03i} + \gamma_3 Z_{ij}$. $X^{(2)}$ was re-scaled by a factor of 10. We further assume that

$$[X_{ij}^{(2)}|Z_{ij}, D_{ij} = 0, X_{ij}^{(1)} = 0] \sim \text{Normal}(\theta_{0ij}, \sigma_0^2), \text{ and}$$
$$[X_{ij}^{(2)}|Z_{ij}, D_{ij} = 1, X_{ij}^{(1)} = 1] \sim \text{Normal}(\theta_{1ij}, \sigma_1^2).$$

It follows that the marginal distribution of $X^{(2)}$ in both the case and the control populations is a mixture of two normal distributions. For each of the parameters β_1 (since Z is a single covariate), β_2, γ_1, γ_2, γ_3 and γ_{01i}, γ_{02i}, γ_{03i}, for $i = 1, 2, \ldots, s$ we use independent Normal(0,10) priors. For both of σ_0^2 and σ_1^2, we use inverse gamma prior with shape parameter 2.9 and scale parameter 0.80. The posterior means of the parameters are taken as our estimates.

Table 4. Results for the benign breast disease data

Method		Education	Regular Medical Checkups	Age at First Pregnancy	CPO
AM	Mean	–0.62	1.47	1.86	–277.11
	s.e.	0.43	0.43	0.50	
	Lower HPD	–1.40	0.65	0.78	
	Upper HPD	0.38	2.42	2.85	
IM	Mean	–0.38	1.61	1.50	–285.86
	s.e.	0.42	0.43	0.52	
	Lower HPD	–1.20	0.78	0.44	
	Upper HPD	0.40	2.53	2.47	
CLR	Mean	–0.41	1.37	1.31	
	s.e.	0.50	0.49	0.56	
	Lower CL	–0.57	0.41	0.21	
	Upper CL	1.39	2.33	2.41	

The results are presented in Table 4. Here, "Mean" is the posterior mean, "s.e." is the posterior standard deviation, "Lower HPD" and "Upper HPD" are the lower and the upper end of the HPD region respectively,"Lower CL"

and "Upper CL" are the lower and the upper end of the confidence limit respectively. For the independence model, we assume that the marginal distribution of $X^{(2)}$ is a single normal distribution. Here, the important exposure age at first pregnancy, had 11% missing values in the original data-set. We notice some interesting numerical differences between AM and IM estimates with AM estimates reflecting a more pronounced effect of the established risk factors. CLR is less efficient than any of the Bayesian methods in the presence of missingness.

3.4. Model comparison. An important part of the above analysis is model comparison. For each of the above examples, we compute the Conditional Predictive Ordinate (CPO) statistic (Gelfand and Dey, 1994) under the AM and IM models (models 1 and 2, say). Let f_1 and f_2 be the two densities under models 1 and 2. Then conditional predictive ordinate (CPO) for the observation j under model i, $i = 1, 2$, is given by

$$CPO_{ij} = f_i(o_j|o_{-j}) = \int f_i(o_j|\theta_i)\pi(\theta_i|o_{-j})d\theta_i,$$

where o_j is the jth observation, o_{-j} is the data without the jth observation and θ_i is the set of parameters under model i. The ratio CPO_{1j}/CPO_{2j} measures how well model 1 supports the observation o_j compared to model 2 based on the remaining data o_{-j}. As an overall aggregate summary of how well the data are supported by model i, one often uses the sum

$$CPO_i = \sum_{j=1}^{n} \log(CPO_{ij}).$$

A better model leads to a higher value of CPO statistic. Note that we can rewrite CPO_{ij} as

$$CPO_{ij} = \left(\int \frac{1}{f_i(o_j|\theta_i)} \pi(\theta_i|\mathbf{o})d\theta_i \right)^{-1},$$

where $\mathbf{o} = (o_1, \ldots, o_n)$. So CPO_{ij} can be estimated (Chen et al., 2000) by

$$CPO_{ij} = \left(\frac{1}{L} \sum_{k=1}^{L} \frac{1}{f_i(o_j|\theta_i^{(k)})} \right)^{-1},$$

where $\theta_i^{(1)}, \ldots, \theta_i^{(L)}$ are L MCMC samples of the parameter(s) θ_i.

For the LA cancer study (Example 1), there is essentially no difference between the CPO values for the two models, whereas for both the colon cancer study and the benign breast disease study (Examples 2 and 3), the AM model has a higher value of the CPO statistic than the IM model, indicating that the AM model is more adequate in those examples.

4 Simulation

To study the effectiveness of association modelling in the context of matched case-control study, we performed a small simulation study. The performance of the association modelling was compared with the independence modelling and the conditional logistic regression method in two different situations, when two exposure variables are associated, and when they are independent.

To simulate a realistic data-set, we used the endometrial cancer data as a prototype. We generated a hypothetical 1:1 matched data-set with 100 strata, a covariate Z and two binary exposure variables $X^{(1)}$ and $X^{(2)}$. The true values of the parameters β_1, β_{21}, β_{22}, and γ_{11}, γ_{12} and γ_{13} were taken as 2.18, 1.11, 0.67, –1.40, 0.19, and 0.34 respectively. These are close to the estimates of the parameters obtained by analysing the data-set through association modelling approach. The stratum specific intercept terms γ_{01i}, γ_{02i}, and γ_{03i} were generated from Normal$(-1.40, 2.37)$, Normal$(0.42, 2.85)$, and Normal$(-1.8, 2.38)$ respectively. The means and the variances of these distributions were chosen in the ballpark of the corresponding posterior distributions in our data analysis with association model.

In the original data-set, 58% of the subjects used estrogen, so the covariate Z was generated as a Bernoulli random variable with success probability 0.58.

Second, we generated the binary disease indicator D. For the i^{th} stratum, it may be easily noted that

$$P(D_{i1} = 1 | D_{i1} + D_{i2} = 1, Z_{i1}, Z_{i2}, S_i) = \frac{1}{1 + \exp\{\beta_1(Z_{i2} - Z_{i1})\} \frac{C_{i2}^{(1)}}{C_{i2}^{(0)}} \frac{C_{i1}^{(0)}}{C_{i1}^{(1)}}}, \tag{4.1}$$

where $C_{ij}^{(0)}$ and $C_{ij}^{(1)}$ are as defined in (3.1) and (3.2). The disease indicators were generated both for the independence model with $\lambda_{ij} = 0$ and for the association model with $\lambda_{ij} = \gamma_{03i} + \gamma_{13} Z_{ij}$. In both cases, $\theta_{ij}^{(1)} = \gamma_{01i} + \gamma_{11} Z_{ij}$,

and $\theta_{ij}^{(2)} = \gamma_{02i} + \gamma_{12} Z_{ij}$. We generated binary random variables with the above probability structure, and conditional on the fact that the simulated value for D_{i1} is a 1 (i.e., the subject is a case), we generated a binary exposure variable ($X_{i1}^{(1)}$) from a Bernoulli distribution with success probability

$$p_{1i} = C_{i1}^{-1} \exp(\theta_{i1}^{(1)} + \beta_{21}) \{1 + \exp(\theta_{i1}^{(2)} + \lambda_{i1} + \beta_{22})\}^{-1}.$$

Conditional on $X_{i1}^{(1)}$, the second binary exposure variable ($X_{i1}^{(2)}$) was generated from a Bernoulli distribution whose success probability is

$$p_{2i|1i} = \frac{\exp\{\lambda_{i1} X_{i1}^{(1)} + \theta_{i1}^{(2)} + \beta_{22}\}}{1 + \exp\{\lambda_{i1} X_{i1}^{(1)} + \theta_{i1}^{(2)} + \beta_{22}\}}.$$

If the simulated value for D_{i1} is 0 (i.e., the subject is a control) we generated $X_{i1}^{(1)}$ from a Bernoulli distribution with success probability $p_{1i} = \exp(\theta_{i1}^{(1)})\{1 + \exp(\theta_{i1}^{(2)} + \lambda_{i1})\}/C_{i1}^{(0)}$. Subsequently, conditional on $X_{i1}^{(1))}$, $X_{i1}^{(2)}$ is generated from a Bernoulli distribution with success probability $p_{2i|1i} = \exp\{\theta_{i1}^{(2)} + \lambda_{i1} X_{i1}^{(1)}\}/(1 + \exp\{\theta_{i1}^{(2)} + \lambda_{i1} X_{i1}^{(1)}\})$. Once we have simulated D_{i1} in the above manner, $D_{i2} = 1 - D_{i1}$. The corresponding exposure variables are then generated accordingly.

We replicated the simulation 100 times, generating 100 different data sets and obtained the parameter estimates for the full data by the three methods (AM, IM and CLR). We then created missing data by randomly deleting 20% of the exposure observations, which essentially generated a prototype MCAR (missing completely at random, in the sense of Little and Rubin, 2002) data. However, as indicated in (2.4), the proposed formulation remains valid for MAR data, as long as the missingness mechanism does not depend on the missing exposure observations. To illustrate this, we also generated MAR data by simulating two indicator variables R_1 and R_2 independently corresponding to two exposure variables with success probability

$$P(R_i = 1 | X_1, X_2, Z, D) = \frac{\exp(0.4 + 2D + 2.5Z)}{1 + \exp(0.4 + 2D + 2.5Z)}, \text{ for } i = 1, 2$$

The parameters of the above probability distribution is so chosen that on an average $P(R = 1)$ is equal to 0.80. For each type of missing data, we recalculated all the estimates by the three methods (i.e., AM, IM, and CLR).

For the Bayesian methods, (i.e., AM and IM), we used Normal$(0, 10)$ prior for all the parameters. We experimented with several choices of priors

and noticed that assuming sharp priors on the nuisance parameters γ_{0i} does affect the posterior distribution of relative risk parameters and recommend using flat priors on the nuisance parameters.

The methods were compared by calculating mean squared error (MSE) of the parameter estimates that take into account both the biases and variances of the estimates.

TABLE 5. RESULTS OF THE SIMULATION STUDY WHEN THE DATA WERE GENERATED WITH ASSOCIATION AMONG EXPOSURES WITH $\lambda_{ij} = \gamma_{03i} + 0.34Z_{ij}$, $\gamma_{03i} \sim \text{NORMAL}(-1.80, 2.38)$

Method		Associated Exposures								
		Full data			MCAR data			MAR data		
		β_1	β_{21}	β_{22}	β_1	β_{21}	β_{22}	β_1	β_{21}	β_{22}
AM	Mean	2.1866	1.0512	0.6353	2.1596	1.0349	0.6369	2.1420	1.0668	0.6559
	Var	0.0737	0.0660	0.0586	0.0797	0.0703	0.0716	0.0709	0.0789	0.0813
	MSE	0.0737	0.0695	0.0598	0.0801	0.0759	0.0726	0.0723	0.0807	0.0814
IM	Mean	2.1582	1.0117	0.6057	2.1026	0.9908	0.5965	2.1239	1.0281	0.6278
	Var	0.0761	0.0661	0.0586	0.0853	0.0814	0.0806	0.0739	0.0881	0.0979
	MSE	0.0766	0.0757	0.0627	0.0913	0.0955	0.0861	0.0771	0.0948	0.0997
CLR	Mean	2.2096	1.1489	0.6866	2.3169	1.1743	0.6625	2.4427	1.1812	0.7087
	Var	0.1167	0.1264	0.0867	0.2975	0.4039	0.2920	0.1455	0.3540	0.1527
	MSE	0.1176	0.1279	0.0869	0.3163	0.4081	0.2921	0.2145	0.3591	0.1543

TABLE 6. RESULTS OF THE SIMULATION STUDY WHEN THE DATA WERE GENERATED WITH INDEPENDENT EXPOSURES WITH LOG-ODDS RATIO $\lambda_{ij} = 0.00$

Method		Independent Exposures								
		Full data			MCAR data			MAR data		
		β_1	β_{21}	β_{22}	β_1	β_{21}	β_{22}	β_1	β_{21}	β_{22}
AM	Mean	2.0886	1.0669	0.6298	2.1192	1.0344	0.6183	2.1156	1.0826	0.5546
	Var	0.0812	0.0490	0.0710	0.0902	0.0543	0.1176	0.0729	0.0749	0.0698
	MSE	0.0895	0.0508	0.0726	0.0939	0.0600	0.1203	0.0761	0.0755	0.0731
IM	Mean	2.0856	1.0579	0.6661	2.1102	1.0662	0.6400	2.1210	1.0855	0.5794
	Var	0.0821	0.0483	0.0701	0.0942	0.0545	0.1142	0.0719	0.0716	0.0656
	MSE	0.0909	0.0571	0.0702	0.0991	0.0564	0.1181	0.0755	0.0723	0.0738
CLR	Mean	2.1759	1.1300	0.6490	2.3215	1.1265	0.7290	2.3981	1.2262	0.5993
	Var	0.1236	0.0922	0.0949	0.3013	0.1627	0.3678	0.2104	0.2723	0.1744
	MSE	0.1236	0.0926	0.0953	0.3213	0.1629	0.3718	0.2580	0.2858	0.1794

The results of the simulation study under different types of missing data are presented in Tables 5 and 6. Here, "Mean" is the mean of the 100 estimates corresponding to the 100 simulated data-sets while MSE is the mean

squared error. The true parameter values are $\beta_1 = 2.18$, $\beta_{21} = 1.11$ and $\beta_{22} = 0.67$. AM and IM stand for association and independence modelling respectively. The study results are fairly clear. When the exposure variables are truly associated (Table 5) and partially missing, AM performs better than IM and CLR in terms of MSE. There is approximately 12-19% reduction in the MSE of the estimators when comparing AM with IM. When the exposure variables are not associated (Table 6), AM and IM perform comparably and both outperform CLR. AM allows the possibility to borrow information on missing values of an exposure variable from the other observed components of the exposure variable via the association structure and also through the completely observed covariate. But, due to independence assumption, IM is not able to extract information on the missing components of an exposure variable from the observed components of the exposure variable, it can only use the completely observed covariate information. Thus, in the presence of association and missingness IM is less efficient. As a practical guideline, if one observes significant association among the exposure variables in a preliminary or similar study by calculating some ad hoc measures such as correlation coefficient or odds ratio, or if there is a pre-existing scientific basis for expecting the exposures to have a moderately strong association, one should proceed to employ AM method for analysing missing data.

5 Concluding Remarks

To summarize the findings of this article, we note that this is the first attempt to account for multivariate association among exposures in missing data situations in a matched case-control study. Through our examples, we find that the association model outperforms the independence model when one has two strongly associated exposures and an informative completely observed covariate. We also present a brief discussion of treatment of missingness and measurement error under the same framework. As mentioned in the introduction, the proposed model presents a way to deal with categorical and continuous exposures simultaneously. Moreover, the work presented here for binary exposures can be extended to a general categorical exposure in a straightforward way. The model also accounts for stratum heterogeneity on exposure distributions through a stratum-specific intercept term while modelling the parameters in the exposure distribution. Thus this paper offers an ensemble of statistical methods for unorthodox data situations in a matched case-control study.

A limitation of the proposed method is that in the absence of missingness, the analysis does not reduce to CLR analysis as there is a stringent parametric structure of the joint distribution of the exposures. This naturally entails some robustness issues, when the posed parametric exposure density model is incorrect. However, this limitation of our method is germane to any method that poses a parametric model for the exposure distribution. We recommend using our method only in the presence of missingness and recognize this resultant model bias versus efficiency dilemma. Rathouz (2003) proposed an alternative approach for handling missing data, which in fact reduces to CLR analysis with completely observed data, but his techniques could not be adapted to our framework.

Acknowledgement. This research was partially supported by NIH Grant R01 85414, and NSF Grant SES-0317589.

References

Armstrong, B.G., Whittemore, A.S. and Howe, G.R. (1989). Analysis of case-control data with covariate measurement error: Application to diet and colon cancer. *Stat. Med.* **8**, 1151–1163.

Barnard, J., McCulloch, R. and Meng, X-L. (2000). modelling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statist. Sinica*, **10**, 1281–1311.

Breslow, N.E. and Cain, K.C. (1988). Logistic regression for two-stage case-control data. *Biometrika* **75**, 11–20.

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

Carroll, R.J., Gail, M.H. and Lubin, J.H. (1993). Case-control studies with errors in covariates. *J. Amer. Statist. Assoc.*, **88**, 185–199.

Carroll, R.J. and Stefanski, L.A. (1994). Measurement error, instrumental variables and corrections for attenuation with applications to meta-analysis. *Stat. Med.*, **13**, 1265–1282.

Chen, M H., Shao, Q M. and Ibrahim, J.G. (2000). *Monte Carlo Methods in Bayesian Computation*, New York: Springer-Verlag.

Ekholm, A., McDonald, J.W. and Smith, P.W.F. (2000). Association models for a multivariate binary response. *Biometrics*, **56**, 712–718.

Gelfand, A.E. and Dey, D.K. (1994). Bayesian model choice: Asymptotics and exact calculations. *J. Roy. Statist. Soc. Ser B*, **56**, 501–514.

Gustafson, P., Le, N. and Vallee, M. (2000). Bayesian analysis of case-control data with imprecise exposure measurements. *Statist. Probab. Lett.*, **47**, 357–363.

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

HOSMER, D.A. and LEMESHOW, S. (2000). *Applied Logistic Regression*. New York: John Wiley.

HUBERMAN, M. and LANGHOLZ, B. (1999). Application of the missing-indicator method in matched case-control studies with incomplete data. *Amer. J. Epidemiology*, **150**, 1340–1345.

LANG, J.B., MCDONALD, J.W. and SMITH, P.W.F. (1999). Association-marginal modelling of multivariate categorical responses: A maximum likelihood approach. *J. Amer. Statist. Assoc.*, **94**, 1161–1171.

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. (2002). *Statistical Analysis with Missing Data*, 2nd Edition. New York: John Wiley.

MILLER, A.B., HOWE, G.R., JAIN, M., CRAIB, K.J.P. and HARRISON, L. (1983). Food items and food groups as risk factors in a case-control study of diet and colo-rectal cancer. *Internat. J. Cancer*, **32**, 155–161.

MUKHERJEE, B., SINHA, S. and GHOSH, M. (2005). Bayesian analysis of case-control studies. *Bayesian Thinking, Modelling, and Computation*, *Handbook of Statistics* **25**. D.K. Dey, and C.R. Rao, eds., North-Holland, Amsterdam, 793–819.

PAIK, M.C. (2004). Nonignorable Missingness in Matched Case-Control Data Analyses. *Biometrics*, **60**, 306–314.

PAIK, M.C. and SACCO, R. (2000). Matched case-control data analyses with missing covariates. *Appl. Stat.*, **49**, 145–156.

PASTIDES, H., KELSEY, J.L., HOLFORD, T.R. and LIVOLSI, V.A. (1985). An epidemiologic study of fibrocystic breast disease with reference to ductal epithelial atypia. *Amer. J. Epidemiology*, **121**, 440–447.

RATHOUZ, P.J. (2003). Likelihood methods for missing covariate data in highly stratified studies. *J. Roy. Statist. Soc. Ser B*, **65**, 711–723.

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.

ROEDER, K., CARROLL, R.J. and LINDSAY, B.G. (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *J. Amer. Statist. Assoc.*, **91**, 722-732.

ROBERT, C.P. and CASELLA, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag, New York.

RUBIN, D.B. and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Amer. Statist. Assoc.*, **81**, 366–374.

SATTEN, G.A. and CARROLL, R.J. (2000). Conditional and unconditional categorical regression models with missing covariates. *Biometrics*, **56**, 384–388.

SATTEN, G.A. and KUPPER, L.L. (1993a). Conditional regression analysis of the exposure-disease odds ratio using known probability-of-exposure values. *Biometrics*, **49**, 429–440.

SATTEN, G.A. and KUPPER, L.L. (1993b). Inferences about exposure-disease associations using probability-of-exposure information. *J. Amer. Statist. Assoc.*, **88**, 200–208.

Schottenfeld, D. and Fraumeni, J.F. (1996). *Cancer Epidemiology and Prevention*. Oxford University Press, Oxford.

Sinha, S., Mukherjee, B., Ghosh, M., Mallick, B.K. and Carroll, R.J. (2005). Semiparametric Bayesian analysis of matched case-control studies with missing exposure. *J. Amer. Statist. Assoc.*, **100**, 591–600.

Zelen, M. and Parker, R.A. (1986). Case-control studies and Bayesian inference. *Stat. Med.*, **5**, 261–269.

Zhao, L.P. and Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642–648.

Samiran Sinha
Department of Statistics
Texas A&M University
College Station, TX 77843, USA
E-mail: sinha@stat.tamu.edu

Bhramar Mukherjee
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109, USA
E-mail: bhramar@umich.edu

Malay Ghosh
Department of Statistics
University of Florida
Gainesville, FL 32611, USA
E-mail: ghoshm@stat.ufl.edu

Paper received November 2005; revised February 2007.