Biometrics 67, 546–558
June 2011

DOI: 10.1111/j.1541-0420.2010.01453.x

Missing Exposure Data in Stereotype Regression Model: Application to Matched Case–Control Study with Disease Subclassification

Jaeil Ahn,[1] Bhramar Mukherjee,[1,*] Stephen B. Gruber,[2] and Samiran Sinha[3]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[2]Department of Epidemiology, Human Genetics and Internal Medicine, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[3]Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.
*email: bhramar@umich.edu

Summary. With advances in modern medicine and clinical diagnosis, case–control data with characterization of finer subtypes of cases are often available. In matched case–control studies, missingness in exposure values often leads to deletion of entire stratum, and thus entails a significant loss in information. When subtypes of cases are treated as categorical outcomes, the data are further stratified and deletion of observations becomes even more expensive in terms of precision of the category-specific odds-ratio parameters, especially using the multinomial logit model. The stereotype regression model for categorical responses lies intermediate between the proportional odds and the multinomial or baseline category logit model. The use of this class of models has been limited as the structure of the model implies certain inferential challenges with nonidentifiability and nonlinearity in the parameters. We illustrate how to handle missing data in matched case–control studies with finer disease subclassification within the cases under a stereotype regression model. We present both Monte Carlo based full Bayesian approach and expectation/conditional maximization algorithm for the estimation of model parameters in the presence of a completely general missingness mechanism. We illustrate our methods by using data from an ongoing matched case–control study of colorectal cancer. Simulation results are presented under various missing data mechanisms and departures from modeling assumptions.

Key words: Conditional likelihood; Nonignorable missingness; Proportional odds; Stages of cancer; Vector generalized linear model.

1. Introduction

In this article, we propose two methods for handling partially missing covariate data in a stereotype regression model while the data are collected through a matched case–control design. The stereotype regression model was proposed by Anderson (1984) for analyzing categorical outcome data by using category-specific scores and maintaining the homogeneous effect of covariates corresponding to each logit. The model stands intermediate between the baseline category logit model and the proportional odds model in terms of model flexibility and parsimony. The model can be adapted to ordered as well as unordered outcome settings whereas the proportional odds model is used only for ordered data. The stereotype model, however, has been less attractive as an alternative to proportional odds model due to computational burden caused by multiplicative structure of the model parameters. Since Anderson's initial paper, there has been only handful of follow-up papers on this class of models. Greenland (1994) proposed a two-step iterative algorithm followed by bootstrap for estimation of model parameters and their standard errors, respectively. Holtbrügge and Schumacher (1991) used an iteratively reweighted least squares algorithm (Green, 1984) to obtain parameter estimates. Recently, Yee and Hastie (2003) considered the stereotype model as a special case of the reduced rank vector generalized linear model (RR-VGLM) and introduced a fitting approach in the R package VGAM (Yee, 2010). Kuss (2006) presented an in-depth overview on the estimation of the parameters of a stereotype model by employing generalized least squares and discussed alternate implementation procedures in standard statistical software. Kuss (2003) considered an illustrative example using the random effects stereotype regression model. Lunt (2004) considered prediction of ordinal outcomes using this model. Ahn et al. (2009) presented Bayesian inference for ordered and unordered stereotype model. Agresti (2010) contains a full discussion of this model (p. 103–117).

Greenland (1994) pointed out an attractive feature of this model in terms of yielding valid inference under retrospective sampling, like in a case–control study. Alternative ordinal models such as the proportional odds or cumulative logit model do not preserve valid inference under outcome stratified sampling (Mukherjee, Liu, and Sinha, 2007; Mukherjee and Liu, 2009). Moreover, for a matched case–control study, the conditional likelihood principle (Breslow and Day, 1980) may be invoked to eliminate stratum-specific nuisance parameters under the stereotype model structure, whereas the proportional odds model is not amenable to this principle (Mukherjee et al., 2008). With advances in detection and diagnosis techniques for cancer, classification information into finer subtypes of cancers/tumors are often available in existing databases. The stereotype model presents an interesting alternative to model association of risk factors with such subtypes rather than just case–control status. The outcome categories or disease subtypes may or may not be ordered in terms of effect of covariates. The stereotype model allows a unique opportunity for testing such ordering restrictions. Thus the model appears to be an appealing tool for analyzing matched case–control data with finer disease subclassification.

Missingness in exposure values is frequently a concern in matched case–control studies. Naive use of the conditional logistic regression (CLR) on complete-case data renders deletion of the entire stratum containing any missing case observations in matched case–control studies. There exists a substantial amount of literature on handling missing data in matched case–control studies (Paik and Sacco, 2000; Satten and Carroll, 2000; Rathouz, Satten, and Carroll, 2002; Rathouz, 2003; Sinha et al., 2005). Depending on the type of missingness mechanism (following the terminology of Little and Rubin, 2002), inference from a naive complete-case CLR analysis may suffer in different ways. If the probability of missingness does not depend on observed data, that is, the data are missing completely at random (MCAR), such analysis will yield consistent but less efficient estimates. If the missingness depends on completely observed data, disease status, or matching variables, that is, if the data are missing at random (MAR), this analysis yields biased and inefficient estimates. All of the above references consider methods to handle MAR data in matched case–control studies.

If the missingness mechanism depends on unobserved exposure values, naive complete-case CLR, as well as the above methods to handle MAR data, can lead to biased and inconsistent results. Paik (2004) used a parametric approach to handle such informative missingness (IM) in matched case–control studies using a pseudo-likelihood. After the first timely investigation of Paik (2004) for handling IM in matched case–control studies, Sinha and Maiti (2008) carried out a comprehensive comparison of Paik's approach with an alternative full-likelihood based approach. Both of these papers use the expectation/maximization (EM) algorithm to estimate model parameters and to derive standard error estimates. None of the above papers, however, consider the problem of modeling disease subclassification, and do not involve the stereotype regression model. Sinha, Mukherjee, and Ghosh (2004) did consider the problem of missing exposure data with multiple disease states using a polytomous regression model but not under IM. The parametric structure of the stereotype model leads to new computational issues and there is no literature on handling missingness under this class of models. In this article, we propose an expectation conditional maximization (ECM) approach and a full Bayesian (FB) approach to handle missing data under the stereotype model. The methods are applied to analyze the association between use of statins (a lipid lowering drug), physical activity, and different stages of colorectal cancer in an ongoing population-based matched case–control study (Poynter et al., 2005).

The rest of the article is organized as follows. In Section 2.1, we introduce the stereotype regression model. In Section 2.2, we describe the conditional likelihood under a matched case–control setting, without any missingness. In Section 2.3, we present the likelihood formulation with partially observed data, with a model for missing data and selection probability mechanism. In Section 3, we discuss the computational strategies to estimate the model parameters, namely, the ECM and the full Bayes strategy. We illustrate our methods via analyzing data from the Molecular Epidemiology of Colorectal Cancer (MECC) Study in Section 4. Finally, we carry out a simulation study to compare properties of the different estimation strategies in terms of bias and mean squared error (MSE) under different missingness mechanisms in Section 5. Section 6 presents brief concluding remarks.

Before concluding this section, we highlight two novel features of this article. To the best of our knowledge, there is no literature on handling missing data under the stereotype model. The current article is also the first one to present a full Bayesian framework to deal with nonignorable missingness in matched case–control studies for binary/categorical outcomes. We compare the performance of both FB and ECM in terms of simulation studies under an array of missingness mechanisms and model misspecification.

2. Models and Assumptions

In this section, we introduce the key ingredients of our likelihood, starting with the stereotype model specification, the complete data likelihood, then followed by models for the selection probability and the distribution of the missing exposure.

2.1 *The Stereotype Regression Model*

The stereotype model is nested within the family of polytomous logistic regression models (Agresti, 2002). The polytomous logistic regression model for a categorical response variable Y with $K+1$ categories and a p-dimensional vector of explanatory variables $\boldsymbol{X}$ is denoted by

$$p(Y=k\,|\,\boldsymbol{X}) = \frac{\exp\left(\beta_{0k} + \boldsymbol{\beta}_k^\top \boldsymbol{X}\right)}{\sum_{l=0}^{K} \exp\left(\beta_{0l} + \boldsymbol{\beta}_l^\top \boldsymbol{X}\right)}, \tag{1}$$

for $k = 0, 1, \ldots, K$ with constraints $\beta_{00} \equiv \boldsymbol{\beta}_0 \equiv 0$. The $p \times 1$ parameter vector $\boldsymbol{\beta}_k$ denotes the log odds ratio of category $Y = k$ relative to baseline category $Y = 0$. Anderson (1984) proposed the stereotype model by imposing a structure on $\boldsymbol{\beta}_k$ such that $\boldsymbol{\beta}_k = \phi_k \boldsymbol{\beta}$. The *stereotype* regression model can thus be represented as

$$p(Y=k\,|\,X) = \frac{\exp(\beta_{0k} + \phi_k \boldsymbol{\beta}^\top \boldsymbol{X})}{\sum_{l=0}^{K} \exp(\beta_{0l} + \phi_l \boldsymbol{\beta}^\top \boldsymbol{X})}, \tag{2}$$

for $k = 0, 1, \ldots, K$. For identifiability of the parameters, we assume $\beta_{00} = \phi_0 \equiv 0$ and $\phi_K \equiv 1$. The number of parameters to be estimated in (2) is $(2K-1)+p$, compared to $K + (p \times K)$ parameters in the polytomous logit model (1). The stereotype model (2) is nested within the class of polytomous logit models (1) and thus the two models can be compared via a likelihood ratio test. Note that the stereotype model reduces to the standard logistic regression model when the outcome is binary, that is, $0 = \phi_0 < \phi_1 = 1$. The stereotype model can be extended to accommodate ordered outcomes with a monotonicity constraint on the category-specific scores, namely, $0 \equiv \phi_0 \leq \phi_1 \leq \cdots \leq \phi_K \equiv 1$. The ordering constraint can be tested in light of the data by comparing the ordered and unordered model using a likelihood ratio test. In contrast, the other popular choice for modeling ordered data, namely, the proportional odds model, is not nested in (1) or (2). The proportional odds model assumes an identical effect of the covariates corresponding to each cumulative probability, reducing the number of parameters to be estimated to $K + p$. The stereotype model allows slightly more flexible structure for the covariate effects when compared to the proportional odds model. One can actually test the indistinguishability of covariate effects on outcome categories k and l by testing $H_0 : \phi_k = \phi_l$ in (2) and potentially collapse categories with similar category-specific scores. However, the limitations of the model are nonlinearity in the parameters due to product terms in $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ and the lack of identifiability of the parameters under the global null hypotheses of $H_0 : \boldsymbol{\beta} = 0$, leading to nonstandard asymptotic theory for likelihood-based inference.

2.2 *Stereotype Regression in Matched Case–Control Studies*

As Greenland (1994) pointed out, the stereotype model leads to consistent estimates of the parameters of interest, namely, $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$, under outcome-stratified sampling. Since asymptotic efficiency of a prospective categorical outcome model with multiplicative intercept structure is established in Scott and Wild (1997) and the stereotype model belongs to this class, asymptotic efficiency results follow under the assumption of a general unconstrained distribution for the exposure vector $\boldsymbol{X}$. Anderson (1984) specifically recommended this model for categorical outcomes that are not generated by segmenting a latent continuous scale, but are summaries of truly discrete multidimensional outcomes. A natural example for such an outcome is stages of cancer that are typically assessed based on multiple diagnostic criteria. For matched case–control studies with finer disease subclassification, the stereotype model provides additional flexibility in terms of eliminating the matched set specific parameters via the conditional likelihood.

We now describe the stereotype regression model for the specific setting of a matched case–control design. Let Y_{ij} denote the disease state corresponding to the jth subject in the ith stratum (or matched set), with S_i denoting variables which contributed explicitly or implicitly to the matching process leading to the ith stratum. The disease states are classified into one of the K distinct categories $1, 2, \ldots, K$, while the reference control group is denoted by $Y_{ij} = 0$. In each of the N strata, we assume there is one case matched with M controls. For ease of notation, we restrict our attention to a single covariate X_{ij} with potential missingness, the results could again be extended to a set of covariates containing missingness in a straightforward way (Sinha et al., 2007). Let $\boldsymbol{Z}_{ij}$ denote the vector of p completely observed covariates $\boldsymbol{Z}_{ij} = [Z_{ij1}, \ldots, Z_{ijp}]^T$ corresponding to the jth subject in the ith stratum. The stratified disease risk model is described as

$$p(Y_{ij} = k \mid X_{ij}, \boldsymbol{Z}_{ij}, S_i) = \frac{\exp\left\{\beta_{0k}(S_i) + \phi_k\left(\beta_1 X_{ij} + \boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij}\right)\right\}}{\displaystyle\sum_{l=0}^{K} \exp\left\{\beta_{0l}(S_i) + \phi_l\left(\beta_1 X_{ij} + \boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij}\right)\right\}}, \quad k = 0, \cdots, K. \tag{3}$$

The $\beta_{0k}(S_i)$ are category-specific intercepts that could vary with strata. For identifiability, $\beta_{00}(S_i) = \phi_0 \equiv 0$ and $\phi_K \equiv 1$. The change in the log odds of an individual being in the kth disease category versus being a control, for each unit increase in X is given by $\phi_k \beta_1$. Without loss of generality, let us assume that the first subject in each stratum is the case and remaining are controls. To eliminate the stratum-specific nuisance parameters $\beta_{0k}(S_i)$, we use the conditional likelihood, by conditioning on the event $\sum_{j=1}^{M+1} Y_{ij} = k_i$, in the ith stratum, where k_i is the observed disease state corresponding to the case subject in the ith stratum, $k_i = 1, \ldots, K$.

Thus the conditional likelihood when we have complete data is given by

$$\begin{aligned} L_c &= \prod_{i=1}^{N} P\left(Y_{i1} = k_i, Y_{i2} = \cdots = Y_{iM+1} = 0 \mid \{X_{ij}, \boldsymbol{Z}_{ij}\}_{j=1}^{M+1}, S_i, \sum_{j=1}^{M+1} Y_{ij} = k_i\right) \\ &= \prod_{i=1}^{N} \frac{\exp\left\{\phi_{k_i}\left(\beta_1 X_{i1} + \boldsymbol{\beta}_2^\top \boldsymbol{Z}_{i1}\right)\right\}}{\displaystyle\sum_{j=1}^{M+1} \exp\left\{\phi_{k_i}\left(\beta_1 X_{ij} + \boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij}\right)\right\}}. \end{aligned} \tag{4}$$

For completely observed data one could proceed with Bayesian inference using the above conditional likelihood treating it as a genuine likelihood and impose prior structure on the parameters ϕ_k, β_1, and $\boldsymbol{\beta}_2$ (Ahn et al., 2009). The justification for using L_c as a basis of Bayesian inference can be found in Rice (2004).

Remark 1. Our results could be directly extended to the setting of a more general $C_i : M_i$ case:control matching ratio. Under such a matching scheme, the conditioning statistic is the vector $\{C_{ki}, k = 1, \ldots, K\}$, where C_{ki} is the number of cases of each subtype k in stratum i, with $\sum_{k=1}^{K} C_{ki} = C_i$. Exact expression for the conditional likelihood under this general case is presented in Web Appendix A.1.

2.3 *Likelihood Formulation under Missingness in Exposure Values*

Let R_{ij} denote the indicator variable assuming the value 1 if X_{ij} is observed and 0 otherwise. The complete joint conditional likelihood we consider as a basis of our inference is given by $L_{cm} = \prod_{i=1}^{N} L_{cm}^i$, where L_{cm}^i, the contribution of the ith stratum to this full data likelihood, can be factored as

$$\begin{aligned} L_{cm}^i &= p\left(\{R_{ij}, X_{ij}, Y_{ij}\}_{j=1}^{M+1} \mid \{\boldsymbol{Z}_{ij}\}_{j=1}^{M+1}, S_i, \sum_{j=1}^{M+1} Y_{ij} = k_i\right) \\ &= \prod_{j=1}^{M+1} \{p(R_{ij} \mid X_{ij}, Y_{ij}, \boldsymbol{Z}_{ij}, S_i) \times f(X_{ij} \mid Y_{ij}, \boldsymbol{Z}_{ij}, S_i)\} \\ &\quad \times p\left(Y_{i1} = k_i, Y_{i2} = \cdots = Y_{iM+1} = 0 \mid \boldsymbol{Z}_{ij}, S_i, \sum_{j=1}^{M+1} Y_{ij} = k_i\right). \end{aligned}$$

Here f denotes the probability distribution function governing the missing data X_{ij}. In order to evaluate this likelihood, we first assume a selection probability model, governed by parameter $\boldsymbol{\delta}$, namely,

$$p(R_{ij} = 1 \mid X_{ij}, Y_{ij}, \boldsymbol{Z}_{ij}, S_i) = H(X_{ij}, Y_{ij}, S_i, \boldsymbol{Z}_{ij}; \boldsymbol{\delta}), \tag{5}$$

where $H(x, y, z, s; \boldsymbol{\delta})$ defines a valid probability mass function for the missingness indicator R. For example, $H(\cdot)$ might be logistic in $(X_{ij}, Y_{ij}, S_i, \boldsymbol{Z}_{ij})$, with $H(u) = \{1 + \exp(-u)\}^{-1}$. However, the results hold for any binary link function and functional specification of the predictors.

We now need to specify a model for $f(X_{ij} \mid Y_{ij}, \boldsymbol{Z}_{ij}, S_i)$. Based on the results of Satten and Kupper (1993) and Satten and Carroll (2000), by specifying a model for $f(X_{ij} \mid Y_{ij} = 0, \boldsymbol{Z}_{ij}, S_i)$ and the prospective disease risk model (3), one can obtain the distribution of X_{ij} in all disease subclasses, namely, $f(X_{ij} \mid Y_{ij} = k, \boldsymbol{Z}_{ij}, S_i), k = 1, \ldots, K$. This well-known result is presented in Web Appendix A.2. The last term in L^i_{cm}, which remains to be expressed as a function of the ingredients of the assumed model components, can be simplified as

$$p\left(Y_{i1} = k_i, Y_{i2} = \cdots = Y_{iM+1} = 0 \mid \boldsymbol{Z}_{ij}, S_i, \sum_{j=1}^{M+1} Y_{ij} = k_i\right)$$
$$= \frac{p(Y_{i1} = k_i \mid \boldsymbol{Z}_{i1}, S_i)/p(Y_{i1} = 0 \mid \boldsymbol{Z}_{i1}, S_i)}{\sum_{j=1}^{M+1} p(Y_{ij} = k_i \mid \boldsymbol{Z}_{ij}, S_i)/p(Y_{ij} = 0 \mid \boldsymbol{Z}_{ij}, S_i)}.$$

The marginal odds (marginalized over X) of the disease $p(Y = k \mid \boldsymbol{Z}, S)/p(Y = 0, \boldsymbol{Z}, S)$ can again be represented in terms of the control distribution for X and the parameters of the disease risk model. The exact representation is in Web Appendix A.2. The marginal likelihood of observed data after integrating with respect to the distribution of the missing exposure is given by $L^{obs}_{cm} = \prod_{i=1}^{N} L^{i,obs}_{cm}$, where

$$\begin{aligned} L^{i,obs}_{cm} = \prod_{j=1}^{M+1} & \{p^{R_{ij}}(R_{ij} = 1 \mid X_{ij}, Y_{ij}, \boldsymbol{Z}_{ij}, S_i) \\ & \times f^{R_{ij}}(X_{ij} \mid Y_{ij}, \boldsymbol{Z}_{ij}, S_i)\} \\ & \times \left\{ \int p^{(1-R_{ij})}(R_{ij} = 0 \mid X_{ij}, Y_{ij}, \boldsymbol{Z}_{ij}, S_i) \right. \\ & \left. \times f^{(1-R_{ij})}(X_{ij} \mid Y_{ij}, \boldsymbol{Z}_{ij}, S_i) dX_{ij} \right\} \\ & \times \frac{p(Y_{i1} = k_i \mid \boldsymbol{Z}_{i1}, S_i)/p(Y_{i1} = 0 \mid \boldsymbol{Z}_{i1}, S_i)}{\sum_{j=1}^{M+1} p(Y_{ij} = k_i \mid \boldsymbol{Z}_{ij}, S_i)/p(Y_{ij} = 0 \mid \boldsymbol{Z}_{ij}, S_i)}. \end{aligned} \tag{6}$$

Instead of Monte Carlo, numeric, or analytic evaluation of the above integrated likelihood followed by maximization procedures, both of our estimation strategies FB and ECM will be based on the following complete data likelihood, $L^{comp}_{cm} = \prod_{i=1}^{N} L^{i,comp}_{cm}$, where

$$\begin{aligned} L^{i,comp}_{cm} = \prod_{j=1}^{M+1} & \{p^{R_{ij}}(R_{ij} = 1 \mid X_{ij}, Y_{ij}, \boldsymbol{Z}_{ij}, S_i) \\ & \times f^{R_{ij}}(X_{ij} \mid Y_{ij}, \boldsymbol{Z}_{ij}, S_i) \\ & \times p^{(1-R_{ij})}(R_{ij} = 0 \mid X_{ij}, Y_{ij}, \boldsymbol{Z}_{ij}, S_i) \\ & \times f^{(1-R_{ij})}(X_{ij} \mid Y_{ij}, \boldsymbol{Z}_{ij}, S_i)\} \\ & \times \frac{p(Y_{i1} = k_i \mid \boldsymbol{Z}_{i1}, S_i)/p(Y_{i1} = 0 \mid \boldsymbol{Z}_{i1}, S_i)}{\sum_{j=1}^{M+1} p(Y_{ij} = k_i \mid \boldsymbol{Z}_{ij}, S_i)/p(Y_{ij} = 0 \mid \boldsymbol{Z}_{ij}, S_i)}. \end{aligned} \tag{7}$$

Remark 2. Note that in our formulation so far, any parametric or semiparametric model can be used for the distribution of X. One could use a class of exponential family models (as in Paik, 2004) or allow it to be more flexible (as in Rathouz, 2003). A flexible semiparametric model for the distribution of X using a Dirichlet process mixture of normals has also been proposed in Mukherjee et al. (2007). In Web Appendix A.3, we consider the general class of exponential family of distributions for X. We present details for two commonly occurring distributions, the Normal and the Binomial distribution, just to provide the reader a sense of how the expressions can be simplified in those instances.

Remark 3. When the missingness mechanism is MAR, $p(R \mid X, Y, \boldsymbol{Z}, S) = p(R \mid Y, \boldsymbol{Z}, S)$ and assuming that $p(R \mid Y, \boldsymbol{Z}, S)$ does not involve any regression parameters of interest, the contribution of that term to the likelihood can be ignored and the above likelihood L^{comp}_{cm} reduces to the likelihood used in Satten and Carroll (2000) and Sinha et al. (2005), by simply removing the two terms in (7) involving the selection probability model.

3. Parameter Estimation and Inference

3.1 *The ECM Approach*

Based on the complete data likelihood L^{comp}_{cm}, we devise an ECM approach to estimate the model parameters. Let $\boldsymbol{\eta}$ denote the parameters governing the assumed control distribution $f(X \mid Z, S, D = 0)$. For example, if we assume that the exposure distribution in controls belongs to an exponential family, that is,

$$\begin{aligned} f(X_{ij} \mid Y_{ij} = 0, \mathbf{Z}_{ij}, S_i) = \exp[\, & \xi_{ij}\{\theta_{ij} X_{ij} - b(\theta_{ij})\} \\ & + c(\xi_{ij}, X_{ij})], \end{aligned}$$

where the canonical parameters θ_{ij} are modeled as a regression function of the completely observed covariates, namely, $\theta_{ij} = \kappa_0 + \boldsymbol{\kappa}_1^\top \mathbf{Z}_{ij} + \kappa_2 S_i$, capturing the dependence of the distribution X on $\boldsymbol{Z}$ and S and ξ_{ij} are the scale parameters. Let, $\boldsymbol{\eta} = (\kappa_0, \boldsymbol{\kappa}_1, \kappa_2, \xi)$. If we denote the entire parameter vector, as $\Theta = (\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\eta}, \boldsymbol{\delta})$, based on Web Appendix A.3, the complete-data log-likelihood, say $l^{comp}_{cm}(\Theta)$ can be obtained by taking log of (7),

$$\begin{aligned}&\underbrace{\sum_{(i,j):R_{ij}=1}\left[\xi_{ij}\left\{\theta^*_{ij}X_{ij}-b\left(\theta^*_{ij}\right)\right\}+c(\xi_{ij},X_{ij})+\log H\left(\delta_0+\delta_1X_{ij}+\delta_2Y_{ij}+\delta_3S_i+\delta_4^\top\boldsymbol{Z}_{ij}\right)\right]}_{L_1(\Theta)}\\&+\underbrace{\sum_{(i,j):R_{ij}=0}\left[\xi_{ij}\left\{\theta^*_{ij}X_{ij}-b\left(\theta^*_{ij}\right)\right\}+c(\xi_{ij},X_{ij})+\log\left\{1-H\left(\delta_0+\delta_1X_{ij}+\delta_2Y_{ij}+\delta_3S_i+\delta_4^\top\boldsymbol{Z}_{ij}\right)\right\}\right]}_{L_2(\Theta)}\qquad(8)\\&+\underbrace{\sum_{i=1}^{N}\left\{\phi_{k_i}\boldsymbol{\beta}_2^\top\boldsymbol{Z}_{i1}+\xi_{i1}\left\{b\left(\theta^*_{i1}\right)-b(\theta_{i1})\right\}+\log\left(\sum_{j=1}^{M+1}\exp\left[\phi_{k_i}\boldsymbol{\beta}_2^\top\boldsymbol{Z}_{ij}+\xi_{ij}\left\{b\left(\theta^*_{ij}\right)-b(\theta_{ij})\right\}\right]\right)\right\}}_{L_3(\Theta)},\end{aligned}$$

where $\theta^*_{ij}=\theta_{ij}+I(Y_{ij}=k_i)\xi_{ij}^{-1}\phi_{k_i}\beta_1$. Using the notations $L_1(\Theta), L_2(\Theta)$, and $L_3(\Theta)$ as defined via (8), we can characterize the *E*-step at the $(t+1)$th iteration of a standard EM algorithm by computing the expectation of $l^{comp}_{cm}(\Theta)$ as,

$$E\left\{l^{comp}_{cm}\left(\Theta^{(t+1)}\right)\right\}=L_1\left(\Theta^{(t+1)}\right)+E\left\{L_2\left(\Theta^{(t+1)}\right)\right\}+L_3\left(\Theta^{(t+1)}\right),\qquad(9)$$

where the expectation E is taken with respect to $f(X_{ij}\,|\,Y_{ij},\boldsymbol{Z}_{ij},S_i,R_{ij}=0,\Theta^{(t)})$ which in turn can be expressed as

$$\begin{aligned}&\frac{p(R_{ij}=0\,|\,X_{ij},Y_{ij},\boldsymbol{Z}_{ij},S_i)f(X_{ij}\,|\,Y_{ij},\boldsymbol{Z}_{ij},S_i)}{p(R_{ij}=0\,|\,Y_{ij},\boldsymbol{Z}_{ij},S_i)}\\&\quad=\frac{p(R_{ij}=0\,|\,X_{ij},Y_{ij},\boldsymbol{Z}_{ij},S_i)f(X_{ij}\,|\,Y_{ij},\boldsymbol{Z}_{ij},S_i)}{\displaystyle\int p(R_{ij}=0\,|\,X_{ij},Y_{ij},\boldsymbol{Z}_{ij},S_i)f(X_{ij}\,|\,Y_{ij},\boldsymbol{Z}_{ij},S_i)dX_{ij}}.\end{aligned}\qquad(10)$$

The integral in (10) is replaced by sum for a discrete exposure X. If we have a standard distributional form for (10), for example, when X is binary, we can obtain an analytical expression for $E\{L_2(\Theta^{(t+1)})\}$. However, Monte Carlo generation or use of other numerical integration routines may be necessary at the *E*-step, depending on the form of the distribution of $f(X_{ij}\,|\,Y_{ij},\boldsymbol{Z}_{ij},S_i)$. In the *M*-step, we maximize (9) at the $(t+1)$th iteration with respect to $\Theta^{(t+1)}$ conditioning on the previously obtained values of $\Theta^{(t)}$.

The above *M*-step may lead to computational complexity with high-dimensional parameter spaces. To handle this difficulty, a modification was proposed by Meng and Rubin (1993) to accelerate the EM algorithm by replacing the *M*-step with a rather simpler conditional maximization (CM) step. With the nonlinearity in ϕ and $\boldsymbol{\beta}$, adopting the ECM is extremely helpful for the stereotype model where the EM often fails to converge. The ECM is in the spirit of Greenland's two-step procedure for stereotype models (Greenland, 1994), where the maximization problem is simplified by iteratively maximizing in terms of $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$. In the $(t+1)$th step of the ECM, we maximize the likelihood in terms of $\boldsymbol{\beta}^{(t+1)}$, for given values of the other parameters obtained at step t, namely, $(\boldsymbol{\phi}^{(t)},\boldsymbol{\eta}^{(t)},\boldsymbol{\delta}^{(t)})$ rather than maximizing the joint likelihood in terms of all parameters $(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\eta},\boldsymbol{\delta})$. Then we maximize the likelihood with respect to $\boldsymbol{\phi}^{(t+1)}$ fixing $(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\eta}^{(t)},\boldsymbol{\delta}^{(t)})$, and continue iteratively. Similar to the EM, we repeat *E*-step and *CM*-step until the convergence condition is met. The conditional maximization is performed via the Nelder–Mead optimization routine.

Remark 4. The standard errors corresponding to the estimated parameters can be obtained by inverting the observed Fisher information as described in Louis (1982):

$$I(\Theta)=-E\left[\frac{\partial^2}{\partial\Theta\partial\Theta^\top}\left\{\log L^{comp}_{cm}(\Theta)\right\}\right]_{\Theta=\hat\Theta}.\qquad(11)$$

We compute the above expectation with respect to the conditional distribution $f(X\,|\,Y,\boldsymbol{Z},S,R=0)$ by Monte Carlo average of the second derivative of the log likelihood. We evaluated each hessian matrix via a numerical approximation in `R` package `hessian:numDeriv`. We evaluate the full hessian matrix at once and do not sequentially condition on remaining parameters, which is known to suffer from invalid standard error estimates (Lall et al., 2002).

3.2 *Bayesian Approach*

Prior specification. The likelihood used for Bayesian inference is again the complete data likelihood in (7). There are four subsets of parameters $(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\eta},\boldsymbol{\delta})$ under consideration. Our main interest lies in $\boldsymbol{\beta}^{(p+1)\times1}=(\beta_1,\boldsymbol{\beta}_2^{p\times1})$ and $\boldsymbol{\phi}^{(K-1)\times1}=(\phi_1,\ldots,\phi_{K-1})$ in the disease risk model (4). The two ancillary sets of parameters involve the $\boldsymbol{\delta}^{(p+4)\times1}=(\delta_0,\delta_1,\delta_2,\delta_3,\boldsymbol{\delta}_4^{p\times1})$ parameters in the selection probability model and the parameters $\boldsymbol{\eta}=(\kappa^{(p+2)\times1},\xi)$, where $\kappa^{(p+2)\times1}=(\kappa_0,\boldsymbol{\kappa}_1^{p\times1},\kappa_2)$, used in modeling the exposure distribution in the control population. To formulate the full conditionals, we assume series of prior distributions on these four sets of parameters.

In this article, we generally consider the following set of mutually independent priors on Θ:

$$\begin{aligned}&\pi(\boldsymbol{\beta})\sim N_{(p+1)}(\boldsymbol{\mu}_\beta,\sigma^2_\beta\boldsymbol{I}),\quad\pi(\boldsymbol{\delta})\sim N_{(p+4)}(\boldsymbol{\mu}_\delta,\sigma^2_\delta\boldsymbol{I}),\\&\pi(\boldsymbol{\kappa})\sim N_{(p+2)}(\boldsymbol{\mu}_\kappa,\sigma^2_\kappa\boldsymbol{I}),\quad\pi(\boldsymbol{\phi})\sim N_{(K-1)}(\boldsymbol{\mu}_\phi,\sigma^2_\phi\boldsymbol{I}).\end{aligned}\qquad(12)$$

On ξ, the scale parameter of the exponential family, we adopt a suitable prior given the specific distribution; for example, we can assume a uniform prior on the logarithmic standard deviation for the distribution of missing data, say, X^m following a Normal distribution. Let us denote by X^o the observed values of X. Based on the complete data likelihood in (7) and the priors described above, we can elicit full conditionals, which are described in detail for specific examples in Web Appendix A.4. and A.5.

Bayesian computation. Following the data augmentation idea of Tanner and Wong (1987) we iterate the following two steps for iteratively generating observations from the joint full conditional of $(\boldsymbol{X}^m, \Theta \,|\, Y, Z, S, \boldsymbol{X}^o)$. At iteration $t+1$,

- (a): Sample $\boldsymbol{X}^m_{(t+1)}$ from density $P(\boldsymbol{X}^m \,|\, \Theta_{(t)}, Y, \boldsymbol{X}^o, \boldsymbol{Z}, S, R)$,
- (b): Sample $\Theta_{(t+1)}$ from density $P(\Theta \,|\, Y, \boldsymbol{X}^o, \boldsymbol{X}^m_{(t+1)}, \boldsymbol{Z}, S, R)$,

where $\Theta^{(t)} = (\boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}^{(t)}, \boldsymbol{\eta}^{(t)},\ \boldsymbol{\delta}^{(t)})$ are obtained at the previous iteration *t*. As Tanner and Wong (1987) pointed out, the first step (a), where we sample $\boldsymbol{X}^m$ from the full conditional distribution, is analogous to "multiple imputation" of filling in the missing data values. Also note that in step (a), we in fact sample X^m from the same full conditional distribution that we use at the *E*-step in ECM as given in (9). In step (b), or the "posterior" step, we generate posterior sample of Θ conditional on augmented data. However, instead of working with a finite number of imputed datasets as in multiple imputation, we iterate this process in our Monte Carlo sampling scheme and continue until stochastic convergence.

Given the full conditionals and employing the above data augmentation step, we use a Gibbs sampler (Geman and Geman, 1984) to generate samples from the full conditional distribution of $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\eta}, \boldsymbol{\delta})$ given the augmented data. Note that though the full conditionals do not often have a standard form, they are log-concave when the distribution of X^m is assumed to belong to a general exponential family. In this case, we use the adaptive rejection sampling or ARS (Gilks and Wild, 1992). For situations when the full conditionals are not log-concave, we adopt the adaptive rejection Metropolis sampling (ARMS; Gilks, Best, and Tan, 1995). For each parameter, we generate 50,000 posterior samples and discard the first 10,000 iterations as "burn-in." In order to reduce the inner-cycle correlation, a thinning of five observations was applied. We monitor convergence of the chains using the diagnostic "potential scale reduction factor" (Gelman and Rubin, 1992) provided in the `R` package `CODA` (Plummer et al., 2009). Finally, the remaining posterior sequences are analyzed for evaluating the Bayesian estimates and highest posterior density (HPD) interval.

4. Example: The Molecular Epidemiology of Colorectal Cancer Study

Colorectal cancer (CRC) is the third most common cancer in the Western world. The Molecular Epidemiology of Colorectal Cancer (MECC) study is a population-based case–control study of patients diagnosed with colorectal cancer in northern Israel between March 31, 1998 and March 31, 2004. Controls were 1:1 matched according to age, sex, and self-reported ethnicity (Jewish vs. non-Jewish). Controls were selected in temporal proximity to the time of diagnosis of the cases. Subjects were interviewed on an array of dietary and behavioral risk factors including levels of physical activity, a family history of colorectal cancer, level of vegetable consumption, and use of medications. Physical activity is known to reduce the risk of CRC by 30–40% according to the informational website of the National Cancer Institute (NCI, 2009). In the MECC dataset, 20% of subjects had missing information on the variable measuring participation in sports or other physical activities. In a high-profile article from the MECC study, Poynter et al. (2005) were the first to point out that the use of statins, a drug used for hypercholesterol, reduces the risk of colorectal cancer (reported OR 0.57, 95% CI: (0.44, 0.73)) after adjusting for other known risk factors, like physical activity, family history of colorectal cancer, the use or nonuse of aspirin or other nonsteroidal anti-inflammatory drugs (NSAID), and level of vegetable consumption. However, no analysis stratified in terms of subtypes of CRC were done in the original study. In the current article, we consider CRC Stage, assigned according to the TNM (Tumor, Node, Metastasis) criteria recommended by American Joint Committee on Cancer (AJCC, 2002) as our categorical outcome ranging from 0 to IV that represents different degree of disease progression. We investigate the effect of physical activity and statin use across CRC stages after adjusting for three other covariates (as mentioned above) via fitting the stereotype model.

We analyzed data on 1784 matched pairs with completely observed data on CRC stage (Y), statin use (Z_1), family history of CRC (Z_2), NSAID use (Z_3), vegetable consumption (Z_4), and partially missing data on physical activity (X). X and (Z_1, Z_2, Z_3) are binary and Z_4 is a trinary covariate (0, 1, 2) indicating low, medium, and high level of vegetable consumption. In our analysis, we consider age, gender, and ethnicity as matching variables $\boldsymbol{S}$ that can affect our selection probability model and the model for control distribution of X. To avoid sparse frequencies, the cancer stage variable Y, was regrouped into four categories 0 (consisting of 1784 controls), 1 (Stage I), 2 (Stage II), and 3 (Stages III and IV). The distribution of subjects in the three case categories were 345 (19.4%), 716 (40.1%), and 723 (40.5%), respectively. The completely observed covariate Z_1 or statin use contained 90% "No" and 10% "Yes." Family history of CRC (Z_2) consists of 90.7% "No" and 9.3% "Yes," while 20% of subjects said "Yes" to NSAID use (Z_3). We consider vegetable consumption (40% Low(0), 30% Medium(1), 30% High(2)) as a continuous covariate. Finally, participation in sports or other physical activity, namely, X contained 29% "No," 51% "Yes," and 20% missing values. Age (S_1) (observed range 19–97) was linearly transformed into a [0, 1] interval. The empirical distribution of transformed age was well approximated by a Normal distribution with mean 0.64 and sd 0.14. For gender (S_2), male is coded as 1 and female as 0, whereas for ethnicity (S_3), Jewish ethnicity is coded as 1 and non-Jewish as 0, with 96% of the subjects/matched pairs coming from Jewish ethnicity.

At the onset, we compared the stereotype model to the polytomous logistic model using only the completely observed data by a number of goodness-of-fit statistics (Web Table 1). The stereotype model indicates better fit in terms of two information criteria used when fitted by maximum likelihood (Kuss, 2006): namely, the Akaike information criterion (AIC) and the Bayes information criterion (BIC). When both models are fitted under a Bayesian framework (Ahn et al., 2009), the stereotype model is preferred by the deviance information criterion (DIC).

We analyzed the MECC data by (a) directly maximizing the conditional likelihood (4) using only the completely observed data (CMLE), (b) the ECM approach, and (c) the full Bayesian method (FB). In order to obtain the CMLE

Table 1

Analysis results for the MECC study data with participation in sports activity X (Yes = 1, No = 0) containing 20% missingness. The set of completely observed covariates are: statin use Z_1 (Yes = 1, No = 0), family history of CRC (Z_2, Yes = 1, No = 0), the use or nonuse of NSAID (Z_3, Yes = 1, No = 0), and the level of vegetable consumption (Z_4, coded as 0, 1, 2 depending on the tertile of consumption, treated as a continuous variable). 1,784 cases are 1:1 matched to controls in terms of age, gender (male = 1, female = 0), and ethnicity (Jewish = 1 vs. non-Jewish = 0). For the CMLE, the conditional likelihood (4) is directly maximized with completely observed data. Under the FB methods the "Est." corresponds to the posterior mean whereas PSD corresponds to posterior standard deviation. For the disease risk parameters, we present 95% Wald confidence intervals (CMLE and ECM) whereas for FB we present 95% highest posterior density (HPD) intervals.

			Method								
		Para-	CMLE			ECM			FB		
Model	Covariates	meter	Est.	SD	(95% CI)	Est.	SD	(95% CI)	Est.	PSD	(95% HPD)
Disease Risk Model	Sports activity	β_1	−0.33	0.10	(−0.52, −0.13)	−0.35	0.08	(−0.50, −0.20)	−0.35	0.08	(−0.52, −0.20)
	Statin use	β_{21}	−0.61	0.16	(−0.97, −0.24)	−0.63	0.15	(−0.92, −0.35)	−0.64	0.15	(−0.93, −0.36)
	Family history of CRC	β_{22}	0.32	0.13	(0.01, 0.62)	0.41	0.12	(0.16, 0.65)	0.41	0.13	(0.17, 0.66)
	NSAID use	β_{23}	−0.34	0.19	(−0.59, −0.09)	−0.45	0.10	(−0.65, −0.26)	−0.46	0.10	(−0.65, −0.26)
	Vegetable intake	β_{24}	−0.27	0.07	(−0.40, −0.13)	−0.21	0.05	(−0.31, −0.12)	−0.22	0.05	(−0.32, −0.13)
Category Specific Score[a]	Stage (I)	ϕ_1	0.78	0.30	(0.19, 1.37)	0.74	0.20	(0.34, 1.14)	0.73	0.22	(0.34, 1.18)
	Stage (II)	ϕ_2	1.26	0.32	(0.63, 1.89)	1.27	0.22	(0.84, 1.70)	1.25	0.23	(0.84,1.72)
Missing Data Model	Intercept	κ_0				−1.15	0.21	(−1.55, −0.76)	−1.16	0.23	(−1.58, −0.69)
	Statin use	κ_{Z_1}				−0.11	0.13	(−0.37, 0.14)	−0.12	0.13	(−0.38, 0.13)
	Family history of CRC	κ_{Z_2}				−0.18	0.09	(−0.36, 0.01)	−0.17	0.10	(−0.36, 0.02)
	NSAID use	κ_{Z_3}				0.34	0.13	(0.09, 0.59)	0.34	0.13	(0.10, 0.62)
	Vegetable intake	κ_{Z_4}				0.42	0.04	(0.34, 0.51)	0.43	0.05	(0.32, 0.51)
	Age	κ_{S_1}				−1.32	0.27	(−1.83, −0.80)	−1.35	0.28	(−1.91, −0.83)
	Gender	κ_{S_2}				0.31	0.07	(0.17, 0.45)	0.31	0.08	(0.16, 0.46)
	Ethnicity	κ_{S_3}				1.17	0.14	(0.90,1.44)	1.18	0.14	(0.92,1.48)
Selection Probability Model	Intercept	δ_0				0.99	0.17	(0.65, 1.33)	1.00	0.22	(0.52, 1.42)
	Sports activity	δ_X				−0.02	0.07	(−0.14, 0.11)	0.00	0.13	(−0.26, 0.25)
	CRC stages	δ_Y				0.04	0.03	(−0.03, 0.10)	0.03	0.03	(−0.03, 0.10)
	Statin use	δ_{Z_1}				0.14	0.15	(−0.15, 0.44)	0.14	0.15	(−0.16, 0.43)
	Family history of CRC	δ_{Z_2}				0.03	0.09	(−0.15, 0.21)	0.03	0.11	(−0.18, 0.25)
	NSAID use	δ_{Z_3}				0.08	0.14	(−0.20, 0.36)	0.13	0.16	(−0.18, 0.43)
	Vegetable intake	δ_{Z_4}				0.19	0.05	(0.09, 0.29)	0.19	0.05	(0.09, 0.29)
	Age	δ_{S_1}				0.00	0.32	(−0.64, 0.64)	0.01	0.30	(−0.58, 0.56)
	Gender	δ_{S_2}				0.02	0.09	(−0.17, 0.19)	0.01	0.09	(−0.16, 0.17)
	Ethnicity	δ_{S_3}				0.18	0.12	(−0.06, 0.43)	0.17	0.14	(−0.10, 0.43)

[a] Other category-specific scores for controls and Stage III, namely, $\phi_0 \equiv 0, \phi_3 \equiv 1$, by the identifiability constraints of a stereotype model.

estimates based on complete data, we used direct maximization of (4) via the Nelder–Mead optimization. Note that using CMLE restricted to completely observed data results in 36% loss of information due to deletion of the entire stratum with any missing covariate. We allowed the missingness data mechanism to potentially depend on $(Y, X, \boldsymbol{Z}, \boldsymbol{S})$ under ECM and FB. For FB, we choose a relatively vague $N(0, 10^4)$ prior on each component of Θ as described in (12). For computing standard errors corresponding to CMLE and ECM, we inverted the observed Fisher information matrix based on complete data and the Monte Carlo evaluated conditional expectation of the Fisher information matrix as specified in (11), respectively. The posterior standard deviations (PSD) for the FB approach were obtained from the standard deviation of the generated posterior sequence.

We present the results of this analysis in Table 1. All three methods produced fairly similar estimates of β_1 and $\boldsymbol{\beta}_2$. The estimated covariate-specific coefficients imply negative association of physical activity, NSAID use, vegetable consumption and use of statins across CRC stages and the effects are highly significant under all methods. Family history increases the estimated risk of CRC. Both FB and ECM have smaller standard errors than the CMLE, due to gain in information by properly using partially observed covariate information. FB and ECM are comparable in terms of the standard errors of the parameter estimates.

Note that the estimated stage-specific parameters $\boldsymbol{\phi}$ are also fairly consistent across three methods. It is evident from the analysis that the association of physical activity and other completely observed covariates with cancer is not homogeneous across different stages of cancer, as the values of ϕ_1 and ϕ_2 differ significantly. A large estimate of ϕ_2, approximately 1.26 from three methods, indicates that the association were more pronounced in Stage 2. The estimates of ϕ_1 and ϕ_2 also imply that there is departure from monotone ordering of the categories in terms of covariate effects, thus the ordered Stereotype model (Anderson, 1984) does not appear to be appropriate for the current analysis. In fact, the posterior probability of the ordering of the categories, that is $p(\phi_0 \equiv 0 < \phi_1 < \phi_2 < \phi_3 \equiv 1 \mid \text{Data})$, was computed from the posterior samples as 0.118, indicating lack of evidence in favor of the ordered stereotype model.

We would like to point out that in the above stereotype model, the log odds-ratio parameters corresponding to each category k as compared to the controls is obtained by the parameters $\phi_k\beta_1$ (for X) and $\phi_k\beta_{2r}$ (for Z_r), $k = 1, 2, 3$, $r = 1, 2, 3, 4$. Bayesian inference has the added advantage of directly generating the posterior of these log odds-ratio parameters directly, instead of resorting to delta theorems and variance approximations that are needed in frequentist inference. Based on the FB analysis, the posterior estimate (95% HPD) of the odds ratios (relative to controls) for physical activity corresponding to categories 1, 2, and 3 are 0.78 (0.66,0.90), 0.65 (0.53,0.78), 0.70 (0.60,0.82), respectively. For use of statins, the corresponding odds ratios are given by 0.63 (0.47,0.83), 0.45 (0.32,0.65), and 0.53 (0.40,0.70), respectively. Figure 1 presents estimated posterior densities of the log odds ratios of each CRC stage versus controls corresponding to physical activity and use of statins, respectively. As pointed out earlier, the nonmonotone trend in the log odds ratios demonstrates that the ordering assumption regarding the category-specific parameters is not tenable for this study. We also tried fitting a proportional odds model to the completely observed data, ignoring the stratification due to matching and the proportional odds assumption was clearly violated with each collapsing of the stage category leading to significantly different estimates for the cumulative relative risk parameter corresponding to each covariate.

Regarding the selection model, the estimated coefficient δ_X is not statistically significant under both ECM and FB (Table 1) but this parameter is only weakly identifiable from observed data and assumed model, thus the test is not very meaningful. Despite certain numerical differences, one can note a general agreement in the point estimates from the complete case analysis compared to estimates under the two models that accommodate missing data in Table 1.

Remark 5. In general, the assumptions regarding the selection probability model (5) are not directly "testable" from the observed data. Thus a sensitivity analysis is required to assess the influence of the modeling assumption on obtained inference. One simple approach toward this is to estimate β_1 and β_{21} under different fixed choices for the coefficients in the selection probability model. To this end, in (5), we fixed δ_X at $(-2,\ 0,\ 2)$ and noted the estimates from FB and ECM. Under FB, $\hat{\beta_1}$ varied from -0.38 to -0.33 and $\hat{\beta_{21}}$ from -0.67 to -0.64. Under ECM, $\hat{\beta_1}$ varies in (-0.35, -0.32) and $\hat{\beta_{21}}$ in (-0.66, -0.63). Similarly, we examined the changes in the parameter estimates of the missing data model, $\hat{\boldsymbol{\kappa}}_Z = (\hat{\kappa}_{Z_1}, \hat{\kappa}_{Z_2}, \hat{\kappa}_{Z_2}, \hat{\kappa}_{Z_4})$, with changes in δ_X. The four components of $\hat{\boldsymbol{\kappa}}_Z$ vary within the range (-0.14, -0.11), (-0.20, -0.16), (0.32, 0.35), and (0.41 0.43), respectively, under FB and (-0.15, -0.08), (-0.18, -0.16), (0.28, 0.34), and (0.32, 0.43), respectively, under ECM. This indicates the impact of changing δ_X on $\hat{\beta}_1, \hat{\beta}_{21}$ and $\hat{\boldsymbol{\kappa}}_Z$ is minimal. Both ECM and FB methods present standard deviations almost identical to the corresponding standard deviations in Table 1 for changing values of δ_X. We carry out more extensive assessment of the robustness properties of our methods via the simulation study in the next section.

5. Simulation Study

We evaluate and compare the performances of the three methods by conducting a small-scale simulation study. The purpose of the simulation study was to assess the methods under various models for the selection probability and the exposure distribution in terms of efficiency and robustness under model misspecification. Mimicking the real data analysis results, we fixed our true parameter values $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\eta}, \boldsymbol{\delta})$ in the range of the point estimates obtained by the three methods. For simplicity, our simulation is based on single Z covariate. We first generate a large cohort of 500,000 subjects, containing information on (Y, X, Z, S). Akin to the statin use variable, we generated Z, from a Bernoulli distribution of success $p = 0.1$. We then generated a potential matching variable S from a Normal(0.6, 0.1^2) distribution, mirroring the age variable in the MECC study. Conditional on Z and S, we generated a binary X from several probability mechanisms as described below in detail. Conditional on X, Z, we generated Y from an unmatched stereotype model. We set the covariate-specific parameters $\beta_1, \beta_2 = (-0.3, -0.7)$ and the category-specific scores as $\boldsymbol{\phi} = (\phi_0, \phi_1, \phi_2) = (0, 0.8, 1.7, 1)$. We selected the three case-category-specific intercepts as $(-1.5,\ -0.5,\ -0.9)$ to make the relative frequency distribution of Y similar to the real data analysis. With this large population base of 500,000 records on $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}$, and $\boldsymbol{S}$, we created a matched case–control dataset in the following way. First, we randomly sampled 1000 cases ($Y \neq 0$) from this large population. Corresponding to each selected case, we chose a matched control randomly from the set of all controls having the value of the matching variable S within 0.05 of the S-value for the selected case. We replicated the aforementioned process 200 times to create 200 matched case–control datasets from this large population under each simulation setting.

Under each simulation configuration, we considered five different schemes of selection probability models. The first four models fall under the class of missingness models we consider

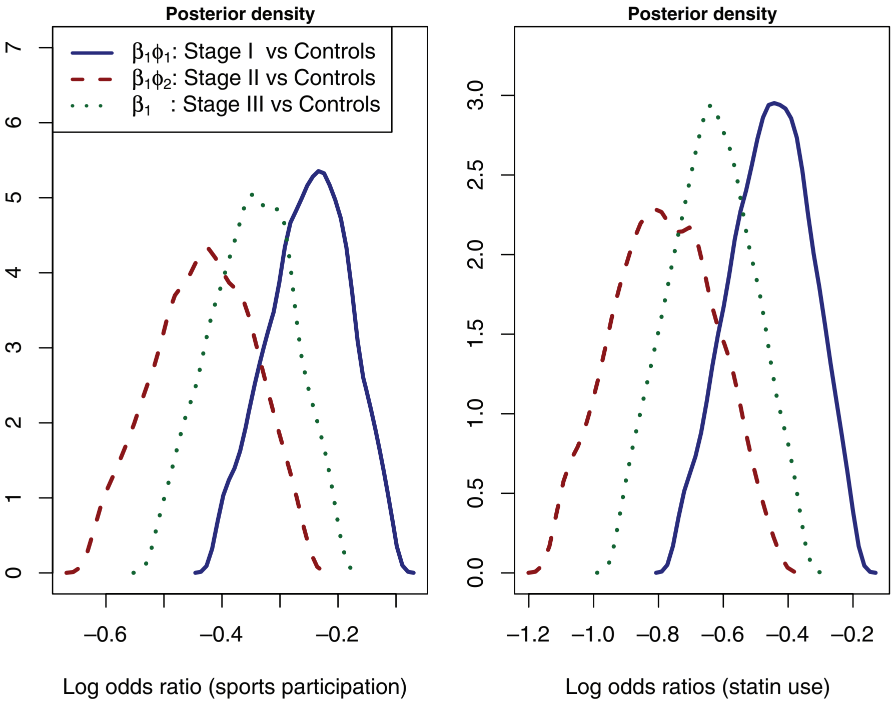

Figure 1. Posterior density plot corresponding to the log odds-ratio parameters in 1:1 matched MECC study data with numerical summaries and estimates as presented in Table 1. The left plot corresponds to participation in sports (X) and the right plot corresponds to statin use (Z_1). The results are based on 10,000 samples generated from the posterior distribution of each parameter. This figure appears in color in the electronic version of this article.

in (5), whereas MM5 involves nonlinear terms in X and Y and violate the modeling assumption of (5).

MM1. Missing completely at random (MCAR): $\text{logit}\{p(R_{ij} = 1 \,|\, Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = 0.8$,

MM2. Missing at random (MAR): $\text{logit}\{p(R_{ij} = 1 \,|\, Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = Y_{ij} + 0.5$,

MM3. Informative missingness (IM): $\text{logit}\{p(R_{ij} = 1 \,|\, Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = X_{ij}$,

MM4. IM: $\text{logit}\{p(R_{ij} = 1 \,|\, Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = 0.5X_{ij} + 0.5Y_{ij} + 0.5$,

MM5. IM: $\text{logit}\{p(R_{ij} = 1 \,|\, Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = X_{ij}Z_{ij} + Y_{ij}X_{ij} + 1$.

The parameters for the above models are chosen in a way to yield the marginal probability of missingness to approximately 20% in each case.

To assess the robustness of our proposed methods under different departures from the assumed model for missing exposure, we consider three scenarios: (a) The exposure model is correctly specified (Table 2); (b) The exposure model is misspecified in terms of a covariate (Table 3); (c) The exposure model is misspecified in terms of a link function (Table 4). Note that since matching is done on the basis of a continuous variable S, the intercept term in the conditional likelihood does not exactly cancel and thus the likelihood is possibly not technically correct in any simulation setting including (a). However, we matched cases and controls within a very narrow interval of S and thus, we do not anticipate any appreciable bias from making this assumption.

Under each simulation setting, we evaluated the performance of three methods: CMLE, ECM, and FB. The corresponding results are presented in terms of the average bias and mean squared errors across the 200 datasets (Tables 2–4). In approximately 3% cases, we failed to obtain estimates from the CMLE approach due to lack of convergence and those simulation iterations are deleted for a fair comparison across the three methods.

Table 2 presents simulation results when the exposure model is correctly specified. We generated exposure $X \,|\, Z, S$ from $H(0.3 + 0.3Z - 1.5S)$. In the presence of noninformative/ignorable missingness (MM1, MM2), the CMLE yields less efficient estimates than the ECM and the FB methods while all three methods are approximately unbiased. With informative missingness and a correctly specified selection model MM4, the ECM and the FB produce less biased estimates than the CMLE in terms of β_1, the coefficient corresponding to X, which is noted to be affected most in the presence of missingness. When model violation exists in terms of the selection probability model having nonlinear product terms XZ and YX (MM5), all three methods produce large biases. Overall, the FB appears to have slightly better mean squared error properties than the ECM.

To assess the effect of model misspecification in the exposure distribution, for example, due to missing a correct covariate term, we introduce a quadratic term S^2, and generate $X \,|\, Z, S$ from $H(0.3 + 0.3Z - 1.5S^2)$ everything else being identical to Table 2 settings. Contrary to our expectation that the full-likelihood based estimates from both FB and ECM will yield enhanced biases compared to the CMLE,

Table 2

Simulation results under correct specification of the exposure model. Here, binary exposure $X \mid Z, S$ is generated from $f(X = 1 \mid Z, S) = H(0.3 + 0.3Z - 1.5S)$. The CMLE, the ECM, and the FB methods are considered. The results are based on 200 simulated datasets, each with 1000 cases and 1000 controls. For each parameter of interest in the disease risk model, we report estimated bias and mean squared error based on the 200 replications. The true values for the parameters of interest are: $\beta_1 = -0.3, \beta_2 = -0.7,\ \phi_1 = 0.8$, and $\phi_2 = 1.7$. Approximately 20% observations in X were missing.

	Method					
	CMLE		ECM		FB	
Parameter	Bias	MSE	Bias	MSE	Bias	MSE
			Complete data			
β_1	0.007	0.009	0.007	0.008	0.030	0.008
β_2	−0.007	0.029	0.002	0.021	0.039	0.021
ϕ_1	−0.053	0.145	−0.072	0.102	−0.036	0.112
ϕ_2	−0.003	0.285	0.003	0.200	0.108	0.223
	MM1. $\text{logit}\{p(R_{ij} = 1 \mid Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = 0.8$					
β_1	−0.023	0.021	−0.015	0.010	0.058	0.011
β_2	−0.046	0.068	−0.027	0.033	0.006	0.031
ϕ_1	0.046	0.301	−0.009	0.132	0.069	0.151
ϕ_2	0.071	0.371	−0.002	0.208	0.112	0.236
	MM2. $\text{logit}\{p(R_{ij} = 1 \mid Y_{ij},\ X_{ij}, Z_{ij}, S_i)\} = Y_{ij} + 0.5$					
β_1	−0.035	0.034	0.033	0.011	−0.010	0.025
β_2	−0.067	0.100	−0.016	0.046	−0.011	0.041
ϕ_1	−0.082	0.307	0.010	0.279	0.026	0.190
ϕ_2	0.070	0.387	0.090	0.293	0.050	0.277
	MM3. $\text{logit}\{p(R_{ij} = 1 \mid Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = X_{ij} + 1$					
β_1	−0.016	0.026	−0.011	0.013	0.006	0.015
β_2	−0.050	0.075	−0.020	0.040	−0.001	0.039
ϕ_1	0.170	0.485	0.118	0.158	0.165	0.175
ϕ_2	0.202	0.647	0.092	0.226	0.171	0.283
	MM4. $\text{logit}\{p(R_{ij} = 1 \mid Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = 0.5X_{ij} + 0.5Y_{ij} + 0.5$					
β_1	−0.115	0.043	−0.036	0.012	−0.051	0.017
β_2	−0.057	0.064	−0.033	0.024	−0.048	0.028
ϕ_1	0.051	0.367	0.050	0.126	0.052	0.099
ϕ_2	0.053	0.686	−0.006	0.150	−0.076	0.245
	MM5. $\text{logit}\{p(R_{ij} = 1 \mid Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = X_{ij}Z_{ij} + Y_{ij}X_{ij} + 1$					
β_1	0.186	0.036	0.170	0.035	0.202	0.046
β_2	−0.126	0.102	0.065	0.043	0.081	0.038
ϕ_1	0.104	0.377	0.046	0.358	0.087	0.259
ϕ_2	0.207	0.969	0.329	0.672	0.342	0.531

which does not make any parametric assumption regarding the exposure distribution, we notice that the results are fairly similar across Table 3 and Table 2 for MM1-MM4 though there is marginally larger bias compared to Table 2. This can be possibly explained by the fact that S^2 and S are not abundantly apart to affect the estimation. Model misspecification in both selection probability and the exposure distribution (MM5), however, results in substantial increase in bias and MSE in the ECM and the FB as shown under MM5.

Table 3

Simulation results under exposure model misspecification in terms of nonlinear predictor in the exposure model. Here, a binary exposure $X \mid Z, S$ is generated under $f(X = 1 \mid Z, S) = H(0.3 + 0.3Z - 1.5S^2)$. The CMLE, the ECM, and the FB methods are considered. The results are based on 200 simulated datasets, each with 1000 cases and 1000 controls. For each parameter, we report estimated bias and mean squared error based on the 200 replications. The true values for the parameters are: $\beta_1 = -0.3, \beta_2 = -0.7, \phi_1 = 0.8$, and $\phi_2 = 1.7$. Approximately 20% observations in X were missing.

	Method					
	CMLE		ECM		FB	
Parameter	Bias	MSE	Bias	MSE	Bias	MSE
			Complete data			
β_1	−0.011	0.011	−0.012	0.010	0.013	0.009
β_2	−0.052	0.032	−0.049	0.031	−0.011	0.029
ϕ_1	0.038	0.126	0.036	0.119	0.080	0.143
ϕ_2	0.010	0.182	0.017	0.172	0.113	0.206
	MM1. $\text{logit}\{p(R_{ij} = 1 \mid Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = 0.8$					
β_1	−0.045	0.023	−0.028	0.017	0.038	0.012
β_2	−0.029	0.068	0.006	0.039	0.023	0.035
ϕ_1	0.047	0.464	0.028	0.261	0.077	0.232
ϕ_2	0.096	0.501	0.090	0.291	0.087	0.290
	MM2. $\text{logit}\{p(R_{ij} = 1 \mid Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = Y_{ij} + 0.5$					
β_1	−0.025	0.021	0.046	0.012	−0.021	0.035
β_2	−0.035	0.054	0.044	0.034	0.040	0.043
ϕ_1	0.044	0.499	−0.039	0.201	0.032	0.119
ϕ_2	0.131	0.503	0.123	0.419	0.063	0.343
	MM3. $\text{logit}\{p(R_{ij} = 1 \mid Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = X_{ij} + 1$					
β_1	−0.005	0.015	−0.016	0.011	0.071	0.013
β_2	−0.008	0.062	−0.018	0.027	0.024	0.026
ϕ_1	0.085	0.262	0.055	0.118	0.120	0.135
ϕ_2	0.194	0.565	0.056	0.172	0.151	0.209
	MM4. $\text{logit}\{p(R_{ij} = 1 \mid Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = 0.5X_{ij} + 0.5Y_{ij} + 0.5$					
β_1	−0.132	0.046	−0.044	0.016	−0.055	0.020
β_2	−0.029	0.060	−0.013	0.037	−0.044	0.028
ϕ_1	−0.052	0.489	0.006	0.124	0.038	0.101
ϕ_2	0.037	0.725	0.070	0.301	−0.052	0.235
	MM5. $\text{logit}\{p(R_{ij} = 1 \mid Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = X_{ij}Z_{ij} + Y_{ij}X_{ij} + 1$					
β_1	0.176	0.048	0.187	0.041	0.222	0.036
β_2	−0.086	0.109	0.047	0.048	0.033	0.041
ϕ_1	0.003	0.570	0.026	0.340	0.102	0.294
ϕ_2	0.243	1.096	0.390	0.867	0.194	0.670

Lastly, we investigate the situation where the link function corresponding to generating $X \mid Z, S$ departs from the logistic link function. Here we generated $X \mid Z, S$ from a mixture of the Burr family of distributions (Burr, 1942),

$$X \mid \boldsymbol{Z}, S \sim \begin{cases} \text{Bernoulli with } f(X = 1 \mid \boldsymbol{Z}, S) \\ \quad = 1 - \{1 + \exp(0.3 + 0.3Z)\}^{-0.7}, \quad S < 0.5 \\ \text{Bernoulli with } f(X = 1 \mid \boldsymbol{Z}, S) \\ \quad = 1 - \{1 + \exp(0.3 + 0.3Z)\}^{-1.3}, \quad S \geq 0.5. \end{cases}$$

Table 4

Simulation results under misspecification in terms of link function corresponding to the exposure distribution. Here, a binary $X \mid Z, S$ is generated from a mixture of Burr family of link functions, $f(X = 1 \mid Z, S) = 1 - \{1 + \exp(0.3 + 0.3Z)\}^{-0.7}$ when $S < 0.5$ and $f(X = 1 \mid Z, S) = 1 - \{1 + \exp(0.3 + 0.3Z)\}^{-1.3}$ otherwise. The CMLE, the ECM, and the FB methods are considered. The results are based on 200 simulated datasets, each with 1000 cases and 1000 controls. For each parameter we report estimated bias and mean squared error based on the 200 replications. The true values for the parameters are: $\beta_1 = -0.3, \beta_2 = -0.7, \phi_1 = 0.8$, and $\phi_2 = 1.7$. Approximately 20% observations in X were missing.

	Method					
	CMLE		ECM		FB	
Parameter	Bias	MSE	Bias	MSE	Bias	MSE
	Complete data					
β_1	−0.008	0.011	−0.021	0.015	0.017	0.011
β_2	−0.036	0.030	0.081	0.019	0.007	0.036
ϕ_1	0.041	0.149	0.098	0.114	0.050	0.142
ϕ_2	0.014	0.273	0.185	0.297	0.131	0.283
	MM1. $\text{logit}\{p(R_{ij} = 1 \mid Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = 0.8$					
β_1	−0.033	0.021	−0.064	0.016	0.045	0.015
β_2	−0.030	0.069	0.058	0.035	0.013	0.034
ϕ_1	0.008	0.328	0.124	0.241	−0.016	0.194
ϕ_2	0.086	0.539	0.112	0.330	0.101	0.299
	MM2. $\text{logit}\{p(R_{ij} = 1 \mid Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = Y_{ij} + 0.5$					
β_1	−0.041	0.023	0.082	0.013	0.089	0.014
β_2	−0.015	0.052	0.076	0.050	0.023	0.039
ϕ_1	−0.025	0.342	0.161	0.301	0.021	0.214
ϕ_2	0.109	0.601	0.148	0.463	0.110	0.387
	MM3. $\text{logit}\{p(R_{ij} = 1 \mid Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = X_{ij} + 1$					
β_1	−0.003	0.021	0.034	0.016	0.042	0.012
β_2	−0.015	0.046	0.015	0.030	0.011	0.024
ϕ_1	0.090	0.324	0.088	0.155	0.091	0.164
ϕ_2	0.210	0.718	0.193	0.277	0.219	0.291
	MM4. $\text{logit}\{p(R_{ij} = 1 \mid Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = 0.5X_{ij} + 0.5Y_{ij} + 0.5$					
β_1	−0.100	0.039	−0.044	0.017	−0.058	0.014
β_2	−0.059	0.059	−0.003	0.031	−0.004	0.026
ϕ_1	0.083	0.321	0.092	0.142	0.091	0.143
ϕ_2	0.056	0.535	0.135	0.322	0.123	0.274
	MM5. $\text{logit}\{p(R_{ij} = 1 \mid Y_{ij}, X_{ij}, Z_{ij}, S_i)\} = X_{ij}Z_{ij} + Y_{ij}X_{ij} + 1$					
β_1	0.201	0.063	0.179	0.072	0.202	0.077
β_2	0.087	0.123	0.074	0.053	0.081	0.061
ϕ_1	−0.036	0.481	−0.071	0.362	0.130	0.351
ϕ_2	−0.341	1.112	0.229	0.803	0.366	0.671

The biases corresponding to the FB and the ECM in Table 4 increase when compared to Tables 2 and 3 with some loss in efficiency. This indicates that this type of link misspecification is possibly more severely affecting the parametric methods of the ECM and the FB than covariate misspecification. Thus the performance of our methods can be dependent upon the nature of the departure from the correct exposure model, producing slightly larger biases than CMLE under MCAR/MAR data (MM1–MM2). However, with IM, both the ECM and the FB lead to improved Bias and MSE properties than the CMLE as the exposure misspecification bias appears to be less, compared to the bias generated by failure to account for nonignorable missingness.

Summarizing our findings, our proposed methods present more efficient estimates than the naive CMLE using completely observed data in the presence of missingness in covariates. In addition, the proposed methods appear to be fairly robust under modest misspecification in the missing exposure distribution. Our approaches do suffer under the incorrect model for informative missingness mechanism. In other more extensive simulation studies under more dramatic departures from the exposure model, we noticed that the ECM approach is less robust than the FB (results are not included). Among the three methods, the FB method has the smallest MSE by virtue of introducing shrinkage effect through prior information. Regarding the secondary model parameters corresponding to the selection probability and the exposure distribution, namely, $\boldsymbol{\delta}$ and $\boldsymbol{\kappa}$, ECM and the FB provide roughly unbiased estimates except for severe model misspecification (MM5 or situation (c)). In order to assess the models in terms of coverage probabilities we compared the coverage of Wald-based confidence intervals of CMLE and ECM with the HPD intervals obtained via FB method (see Web Table 2). We noticed the same phenomena that the ECM and FB have close to nominal coverage probabilities unless there is acute violation in specifying the selection probability model (MM5) while complete-case CMLE suffers when there is nonignorable missingness depending on both X, Y, Z. Finally, we would also like to point out that the computing time needed for the ECM is substantially less than the FB method.

6. Discussion

This article presents a comprehensive approach to handle nonignorable missingness in covariates under the stereotype regression model. Though we focus on matched case–control studies with finer disease subclassification as our primary example, the methods can be adapted to prospective analysis of categorical response data with ordered or unordered response categories using the stereotype class of link functions. We develop an expectation/conditional maximization algorithm as well as a full Bayes procedure with data augmentation and compare these approaches with naive use of conditional maximum likelihood based on complete data. Our real data analysis as well as simulation study establish the methods lead to substantial gain in efficiency compared to the CMLE and are fairly robust under modest departures from the model for missing exposure. However, the methods could perform poorly if the selection probability model is grossly misspecified.

Inference under the stereotype model is burdened with computational and analytical challenges due to embedded nonlinearity and the lack of identifiability in the parametric structure. Missingness further compounds the complexity. The Bayesian paradigm offers flexible alternative modeling approaches and inferential solutions for this class of models. For matched case–control data, the model has an added distinction of accommodating highly stratified data via conditioning and preserving prospective–retrospective conversion of the parameters of interest. The current article is the first attempt to handle a general form of missingness in this class of models. Future research involves considering a more flexible semiparametric model for the exposure distribution, for the missingness mechanism and considering missingness with correlated or clustered observations as in a longitudinal cohort study under the stereotype model. A random effects approach on the stratum effects, instead of using conditional likelihood, is also a plausible alternative and will reduce the bias under data missing at random for complete-case analysis.

7. Supplementary Materials

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

Acknowledgements

The research of Bhramar Mukherjee was partially supported by NSF DMS 07-06935 and NIH grant R03 CA130045-01. The Molecular Epidemiology of Colorectal Cancer Study is supported via NIH grant R01 CA81488. The authors would like to thank Alan Agresti for many helpful discussions regarding the stereotype model. Web Appendices referred throughout this article and the source code for the three methods are available at link `http://www-personal.umich.edu/∼jaeil/`.

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd edition. New York: John Wiley and Sons.

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*, 2nd edition. Hoboken, New Jersey: Wiley.

Ahn, J., Mukherjee, B., Banerjee, M., and Cooney, K. A. (2009). Bayesian inference for the stereotype regression model: Application to a case-control study of prostate cancer. *Statistics in Medicine*, **28,** 3134–3157.

American Joint Committee on Cancer (2002). *AJCC Cancer Staging Manual*, 6th edition, 113–124. New York: Springer.

Anderson, J. A. (1984). Regression and ordered categorical variable. *Journal of the Royal Statistical Society, Series B* **46,** 1–30.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research: Vol. 1—The Analysis of Case-Control Studies.* Lyon, France: IARC Scientific Publications.

Burr, I. (1942). Cumulative frequency functions. *Annals of Mathematical Statistics* **13,** 215–232.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7,** 457–472.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6,** 721–741.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41,** 337–348.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling. *Applied Statistics* **44,** 455–472.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternative (with discussion). *Journal of the Royal Statistical Society, Series B* **46,** 149–192.

Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statistics in Medicine* **13,** 1665–1677.

Holtbrügge, W. and Schumacher, M. (1991). A comparison of regression models for the analysis of ordered categorical data. *Applied Statistics* **40,** 249–259.

Kuss, O. (2003). Modelling physicians' recommendations for optimal medical care by random effects stereotype regression. *Proceedings of the 18th Workshop on Statistical Modelling*, G. Verbeke, G. Molenberghs, M. Aerts, and S. Fieuws (eds), 245–249. Leuven: Katholieke Universiteit Leuven.

Kuss, O. (2006). On the estimation of the stereotype regression model. *Computational Statistics and Data Analysis* **50,** 1877–1890.

Lall, R., Campbell, M. J., Walters, S. J., Morgan, K., and MRC CFAS Co-operative. (2002). A review of ordinal regression models applied on health-related quality of life assessments. *Statistical Methods in Medical Research* **11,** 49–67.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data,* 2nd edition. New York: Wiley.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44,** 226–233.

Lunt, M. (2004). Prediction of ordinal outcomes when the association between predictors and outcome differs between outcome levels. *Statistics in Medicine* **24,** 1357–1369.

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80,** 267–278.

Mukherjee, B. and Liu, I. (2009). A characterization of bias for fitting multivariate generalized linear models under choice-based sampling. *Journal of Multivariate Analysis* **100,** 459–472.

Mukherjee, B., Liu, I., and Sinha, S. (2007). Analysis of matched case-control data with ordinal disease states: Possible choices and comparisons. *Statistics in Medicine* **26,** 3240–3257.

Mukherjee, B., Ahn, J., Liu, I., Rathouz, P. J., and Sanchez, B. N. (2008). On elimination of nuisance parameters in a stratified proportional odds model by amalgamating conditional likelihoods. *Statistics in Medicine* **27,** 4950–4971.

National Cancer Institute. (2009). Physical activity and cancer. Available at: `http://www.cancer.gov/cancertopics/factsheet/prevention/physicalactivity`, accessed 1 June 2009.

Paik, M. C. (2004). Nonignorable missingness in matched case-control data analyses. *Biometrics* **60,** 306–314.

Paik, M. C. and Sacco, R. L. (2000). Matched case-control data analyses with missing covariate. *Journal of the Royal Statistical Society, Series C* **49,** 145–156.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2009). Package `CODA`, Version 0.13-4, Output analysis and diagnostics for MCMC. Avaiable at: `http://cran.r-project.org/web/packages/coda/` accessed 1 May 2009.

Poynter, J. N., Gruber, S. B., Higgins, P. D. R., Almog, R., Bonner, J. D., Rennert, H. S., Low, M., Greenson, J. K., and Rennert, G. (2005). Statins and the risk of colorectal cancer. *The New England Journal of Medicine* **352,** 2184–2192.

Rathouz, P. J. (2003). Likelihood methods for missing covariate data in highly stratified studies. *Journal of the Royal Statistical Society, Series 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.

Rice, K. M. (2004). Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, with applications. *Journal of the American Statistical Association* **99,** 510–522.

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

Satten, G. A. and Kupper, L. (1993). Inferences about exposure-disease associations using probability of exposure information. *Journal of the American Statistical Association* **88,** 200–208.

Scott, A. J. and Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84,** 57–71.

Sinha, S. and Maiti, T. (2008). Analysis of matched case-control data in presence of nonignorable missing exposure. *Biometrics* **64,** 106–114.

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

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

Sinha, S., Mukherjee, B., and Ghosh, M. (2007). Modelling association among bivariate exposures in matched case–control studies. *Sankhya* **69,** 379–404.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82,** 528–550.

Yee, T. W. (2010). The `VGAM` package for categorical data analysis. *Journal of Statistical Software*, **32,** 1–34.

Yee, T. W. and Hastie, T. J. (2003). Reduced-rank vector generalized linear models. *Statistical Modeling* **3,** 15–41.

Received September 2009. *Revised April* 2010.
Accepted April 2010.