Biometrika (2011), **98**, 3, *pp.* 647–662 doi: 10.1093/biomet/asr035

Marginal methods for correlated binary data with misclassified responses

By ZHIJIAN CHEN, GRACE Y. YI and CHANGBAO WU

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

z11chen@uwaterloo.ca yyi@uwaterloo.ca cbwu@uwaterloo.ca

Summary

Misclassification is a longstanding concern in medical research. Although there has been much research concerning error-prone covariates, relatively little work has been directed to problems with response variables subject to error. In this paper we focus on misclassification in clustered or longitudinal outcomes. We propose marginal analysis methods to handle binary responses which are subject to misclassification. The proposed methods have several appealing features, including simultaneous inference for both marginal mean and association parameters, and they can handle misclassified responses for a number of practical scenarios, such as the case with a validation subsample or replicates. Furthermore, the proposed methods are robust to model misspecification in a sense that no full distributional assumptions are required. Numerical studies demonstrate satisfactory performance of the proposed methods under a variety of settings.

Some key words: Misclassification; Odds ratio; Replicate; Unbiased estimating equation; Validation subsample.

1. Introduction

Correlated data, including longitudinal and clustered data, are common in many fields, including epidemiological research and clinical trials. Various models have been developed for analysis of such data, and a wide variety of estimation techniques have been proposed. In contrast to conditional models such as transition models and mixed effects models, marginal models have been widely used in analysing longitudinal or clustered data. A compelling feature of such methods lies in their minimal model assumptions. For example, generalized estimating equations, proposed by Liang & Zeger (1986), focus on estimation of mean parameters, with association parameters between outcomes treated as nuisance. Extensions of this approach can be found, for instance, in Miller et al. (1993) and Molenberghs & Lesaffre (1999), amongst others.

In many epidemiological studies, association structures among repeated outcomes are of scientific interest. For example, understanding the correlation of disease status among family members is often of primary interest in familial studies. Prentice (1988), Carey et al. (1993) and Yi & Cook (2002) extended the generalized estimating equations approach to estimate association parameters for binary data, by specifying a second set of estimating equations. Those methods are useful for simultaneous inference about the mean and association parameters. The validity of these methods requires a critical condition: variables must be precisely measured. However, this requirement is often violated in practice. Misclassification commonly arises with categorical data. Neuhaus (1999, 2002) demonstrated that a naive analysis with misclassification ignored often leads to biased results. With binomial regression models and generalized mixed models Paulino et al. (2003, 2005) proposed Bayesian approaches to handling misclassification in binary data. Cook et al. (2000) described a latent Markov model for longitudinal binary data, whereas

Rosychuk & Thompson (2001, 2003) considered two-state Markov models with misclassified responses.

Relative to a large body of methods on covariate measurement error, research on error-contaminated outcomes, such as misclassified responses, has been quite limited (Carroll et al., 2006), especially under the marginal analysis framework. In this article, we consider marginal regression models for correlated binary data subject to misclassification. We propose estimating equation methods that can correct for misclassification effects under a variety of practical settings. The proposed methods have several appealing features. They accommodate simultaneous inference for both marginal mean and association parameters, and are robust to model misspecification because no full distributional assumptions are required.

2. NOTATION AND MODEL FORMULATION

2·1. *The response process*

Let Y_{ij} be the binary response for the jth subject in cluster i and X_{ij} be the corresponding covariate vector ($i = 1, \ldots, n$; $j = 1, \ldots, m_i$), where n is the number of clusters, and m_i is the number of subjects in cluster i. Denote $Y_i = (Y_{i1}, \ldots, Y_{im_i})^{\mathrm{T}}$ and $X_i = (X_{i1}^{\mathrm{T}}, \ldots, X_{im_i}^{\mathrm{T}})^{\mathrm{T}}$. Let $\mu_{ij} = E(Y_{ij} \mid X_i)$ be the marginal mean of the response, and write $\mu_i = (\mu_{i1}, \ldots, \mu_{im_i})^{\mathrm{T}}$. A generalized regression model is used to link μ_{ij} to the covariates, where $E(Y_{ij} \mid X_i) = E(Y_{ij} \mid X_{ij})$ is assumed (Pepe & Anderson, 1994). That is,

$$g(\mu_{ij}) = X_{ij}^{\mathrm{T}}\beta,$$

where β is a vector of regression parameters, and $g(\cdot)$ is a monotone link function. Typical choices of $g(\cdot)$ include logit, probit and complementary log–log functions. The variance of the response Y_{ij} is specified as $v_{ij} = \mathrm{var}(Y_{ij} \mid X_i) = \mu_{ij}(1 - \mu_{ij})$ accordingly.

When the mean parameters are of primary interest and association parameters are treated as nuisance, the approach of Liang & Zeger (1986) is well suited for parameter estimation. However, to facilitate inference for association parameters that are often of interest for clustered data analysis, one needs to derive a second set of estimating functions to feature association structures. Here we assume that Y_{ij} and $Y_{i'j'}$ are independent when $i \neq i'$, but Y_{ij} and $Y_{ij'}$ may be correlated for $j \neq j'$. Let $Z_{ijj'} = Y_{ij}Y_{ij'}$, $Z_i = (Z_{ijj'}, j < j')^{\mathrm{T}}$, $\mu_{ijj'} = E(Z_{ijj'} \mid X_i)$, and $\xi_i = (\mu_{ijj'}, j < j')^{\mathrm{T}}$.

Odds ratios are often used to facilitate association among correlated binary data (e.g., Lipsitz et al., 1991). For $j < j'$, the odds ratio for Y_{ij} and $Y_{ij'}$ is defined as

$$\psi_{ijj'} = \frac{\mathrm{pr}(Y_{ij} = 1, Y_{ij'} = 1 \mid X_i)\mathrm{pr}(Y_{ij} = 0, Y_{ij'} = 0 \mid X_i)}{\mathrm{pr}(Y_{ij} = 1, Y_{ij'} = 0 \mid X_i)\mathrm{pr}(Y_{ij} = 0, Y_{ij'} = 1 \mid X_i)}.$$

It is often assumed that $\mathrm{pr}(Y_{ij} = y_{ij}, Y_{ij'} = y_{ij'} \mid X_i) = \mathrm{pr}(Y_{ij} = y_{ij}, Y_{ij'} = y_{ij'} \mid X_{ij}, X_{ij'})$. The odds ratios are customarily modelled as

$$\log \psi_{ijj'} = u_{ijj'}^{\mathrm{T}}\alpha,$$

where $u_{ijj'}$ is a set of pair-specific covariates featuring various association structures such as autoregressive or exchangeable structure between Y_{ij} and $Y_{ij'}$. The relationship between $\mu_{ijj'}$ and $\psi_{ijj'}$ is (Lipsitz et al., 1991; Yi & Cook, 2002)

$$\mu_{ijj'} = \begin{cases} [a_{ijj'} - \{a_{ijj'}^2 - 4(\psi_{ijj'} - 1)\psi_{ijj'}\mu_{ij}\mu_{ij'}\}^{1/2}]/\{2(\psi_{ijj'} - 1)\} & (\psi_{ijj'} \neq 1), \\ \mu_{ij}\mu_{ij'} & (\psi_{ijj'} = 1), \end{cases}$$

where $a_{ijj'} = 1 - (1 - \psi_{ijj'})(\mu_{ij} + \mu_{ij'})$.

2·2. *The misclassification process*

When the response Y_{ij} is subject to misclassification, a surrogate version for Y_{ij}, say, S_{ij}, is observed. Let $H_{ij} = \mathrm{I}(S_{ij} = Y_{ij})$ be the indicator variable for misclassification, where $\mathrm{I}(\cdot)$ is the indicator function. Let $H_i = (H_{i1}, \ldots, H_{im_i})^{\mathrm{T}}$, and $S_i = (S_{i1}, \ldots, S_{im_i})^{\mathrm{T}}$. The marginal probability of misclassifying Y_{ij} is assumed to depend only on the jth outcome, given the covariates in cluster i, i.e., $\mathrm{pr}(S_{ij} = 1 \mid Y_i, X_i) = \mathrm{pr}(S_{ij} = 1 \mid Y_{ij}, X_i)$. Let $\tau_{0ij} = \mathrm{pr}(H_{ij} = 1 \mid Y_{ij} = 0, X_i)$ and $\tau_{1ij} = \mathrm{pr}(H_{ij} = 1 \mid Y_{ij} = 1, X_i)$. Alternatively, if we define $\tau_{ij}(y_{ij}) = \mathrm{pr}(H_{ij} = 1 \mid Y_{ij} = y_{ij}, X_i)$, then $\tau_{ij}(y_{ij}) = (1 - y_{ij})\tau_{0ij} + y_{ij}\tau_{1ij}$.

Logistic models may be employed to characterize these probabilities:

$$\mathrm{logit}(\tau_{0ij}) = L_{ij}^{\mathrm{T}}\gamma_0, \quad \mathrm{logit}(\tau_{1ij}) = L_{ij}^{\mathrm{T}}\gamma_1, \tag{1}$$

where γ_0 and γ_1 are vectors of associated regression parameters, and L_{ij} is a set of covariates that reflects various misclassification mechanisms. Let $\gamma = (\gamma_0^{\mathrm{T}}, \gamma_1^{\mathrm{T}})^{\mathrm{T}}$. Covariates L_{ij} may be specified as various forms to feature different misclassification processes. In some situations, L_{ij} is the entire covariate vector X_{ij}; while in extreme cases, L_{ij} can be constant 1, that is, two parameters γ_0 and γ_1 are sufficient to describe the misclassification mechanism. The latter scenario corresponds to a homogeneous misclassification across all observations and clusters, with misclassification independent of covariates and the other outcomes: $\tau_{0ij} = \tau_0 = \mathrm{expit}(\gamma_0)$, and $\tau_{1ij} = \tau_1 = \mathrm{expit}(\gamma_1)$, where $\mathrm{expit}(u) = \exp(u)/\{1 + \exp(u)\}$.

To describe possible dependence between H_{ij} and $H_{ij'}$, we invoke the odds ratios

$$\lambda_{ijj'}(y_{ij}, y_{ij'}) = \frac{\mathrm{pr}(H_{ij} = 1, H_{ij'} = 1 \mid Y_i = y_i, X_i)}{\mathrm{pr}(H_{ij} = 1, H_{ij'} = 0 \mid Y_i = y_i, X_i)} \times \frac{\mathrm{pr}(H_{ij} = 0, H_{ij'} = 0 \mid Y_i = y_i, X_i)}{\mathrm{pr}(H_{ij} = 0, H_{ij'} = 1 \mid Y_i = y_i, X_i)},$$

where it is assumed that $\mathrm{pr}(H_{ij} = h_{ij}, H_{ij'} = h_{ij'} \mid Y_i = y_i, X_i) = \mathrm{pr}(H_{ij} = h_{ij}, H_{ij'} = h_{ij'} \mid Y_{ij} = y_{ij}, Y_{ij'} = y_{ij'}, X_i)$. The odds ratio $\lambda_{ijj'}(y_{ij}, y_{ij'})$ can be modelled by

$$\log\{\lambda_{ijj'}(y_{ij}, y_{ij'})\} = q_{ijj'}^{\mathrm{T}}\nu_{y_{ij}, y_{ij'}},$$

where $q_{ijj'}$ is a vector of covariates that features various types of dependence, and $\nu_{y_{ij}, y_{ij'}}$ is a vector of regression coefficients that may vary with the values of y_{ij} and $y_{ij'}$. Let $\nu = (\nu_{11}^{\mathrm{T}}, \nu_{10}^{\mathrm{T}}, \nu_{01}^{\mathrm{T}}, \nu_{00}^{\mathrm{T}})^{\mathrm{T}}$, and $\eta = (\gamma^{\mathrm{T}}, \nu^{\mathrm{T}})^{\mathrm{T}}$.

For $j < j'$, let $C_{ijj'} = H_{ij}H_{ij'}$, $C_i = (C_{ijj'}, j < j')^{\mathrm{T}}$, $\zeta_{ijj'}(y_{ij}, y_{ij'}) = E(C_{ijj'} \mid Y_{ij} = y_{ij}, Y_{ij'} = y_{ij'}, X_i)$, and $\zeta_i = E(C_i \mid Y_i, X_i)$. Again, $E(C_{ijj'} \mid Y_i, X_i) = E(C_{ijj'} \mid Y_{ij} = y_{ij}, Y_{ij'} = y_{ij'}, X_i)$ is assumed. The relationship between $\zeta_{ijj'}(y_{ij}, y_{ij'})$ and $\lambda_{ijj'}(y_{ij}, y_{ij'})$ is

$$\zeta_{ijj'}(y_{ij}, y_{ij'}) = \begin{cases} \dfrac{b_{ijj'}(y_{ij}, y_{ij'}) - g_{ijj'}(y_{ij}, y_{ij'})}{2\{\lambda_{ijj'}(y_{ij}, y_{ij'}) - 1\}} & \{\lambda_{ijj'}(y_{ij}, y_{ij'}) \neq 1\}, \\ \tau_{ij}(y_{ij})\tau_{ij'}(y_{ij'}) & \{\lambda_{ijj'}(y_{ij}, y_{ij'}) = 1\}, \end{cases}$$

where $b_{ijj'}(y_{ij}, y_{ij'}) = 1 - \{1 - \lambda_{ijj'}(y_{ij}, y_{ij'})\}\{\tau_{ij}(y_{ij}) + \tau_{ij'}(y_{ij'})\}$, and $g_{ijj'}(y_{ij}, y_{ij'}) = [b_{ijj'}^2(y_{ij}, y_{ij'}) - 4\{\lambda_{ijj'}(y_{ij}, y_{ij'}) - 1\}\lambda_{ijj'}(y_{ij}, y_{ij'})\tau_{ij}(y_{ij})\tau_{ij'}(y_{ij'})]^{1/2}$.

Let $\mu_{ijj'}^S = E(S_{ij}S_{ij'} \mid X_i)$ be the marginal mean of $S_{ij}S_{ij'}$ given covariates. It can be shown that $\mu_{ijj'}^S \neq \mu_{ijj'}$. Even under the independence assumption such as $\mathrm{pr}(S_{ij} = s_{ij}, S_{ij'} = s_{ij'} \mid Y_i, X_i) = \mathrm{pr}(S_{ij} = s_{ij} \mid Y_i, X_i)\mathrm{pr}(S_{ij'} = s_{ij'} \mid Y_i, X_i)$, $\mu_{ijj'}^S$ is not equal to $\mu_{ijj'}$. As a consequence, replacing Y_{ij} with S_{ij} in the marginal analysis often leads to biased inference.

3. ESTIMATING EQUATIONS

3·1. *Estimating equations under the true model*

Let $\theta = (\beta^{\mathrm{T}}, \alpha^{\mathrm{T}})^{\mathrm{T}}$ be the vector of response parameters, $D_{1i} = \partial \mu_i^{\mathrm{T}}/\partial\beta$, and $B_{1i} = \mathrm{diag}(v_{i1}, \ldots, v_{im_i})$. When the response variable is free of misclassification, estimates of mean parameters β can be obtained by solving first-order estimating equations (Liang & Zeger, 1986)

$$\sum_{i=1}^{n} U_{1i}(\theta) = 0, \tag{2}$$

where $U_{1i}(\theta) = D_{1i} V_{1i}^{-1} \epsilon_{1i}$, $\epsilon_{1i} = Y_i - \mu_i$, $V_{1i} = \mathrm{cov}(Y_i) = B_{1i}^{1/2} R_{1i}(\theta) B_{1i}^{1/2}$, and $R_{1i}(\theta)$ is the correlation matrix of Y_i with off-diagonal entries $\rho_{ijj'} = (\mu_{ijj'} - \mu_{ij}\mu_{ij'}) \times [\{\mu_{ij}(1-\mu_{ij})\}^{1/2} \{\mu_{ij'}(1-\mu_{ij'})\}^{1/2}]^{-1}$.

Let $D_{2i} = \partial \xi_i^{\mathrm{T}}/\partial\alpha$. Then the second-order estimating equations (Prentice, 1988) for association parameters α can be written as

$$\sum_{i=1}^{n} U_{2i}(\theta) = 0, \tag{3}$$

where $U_{2i}(\theta) = D_{2i} V_{2i}^{-1} \epsilon_{2i}$, and $\epsilon_{2i} = Z_i - \xi_i$. Here V_{2i} is a working covariance matrix for Z_i which is commonly taken as an independence matrix $V_{2i} = \mathrm{diag}\{\mu_{ijj'}(1-\mu_{ijj'}); j < j'\}$. Choosing an independence working matrix might incur some efficiency loss, but it has the appeal of not requiring additional model assumptions for third and fourth moments of the response variables. Moreover, this treatment still retains the unbiasedness of estimating functions $U_{1i}(\theta)$ and $U_{2i}(\theta)$, which ensures a consistent estimator of θ under regularity conditions, e.g., Prentice (1988); Yi & Cook (2002). Let $U_i(\theta) = \{U_{1i}^{\mathrm{T}}(\theta), U_{2i}^{\mathrm{T}}(\theta)\}^{\mathrm{T}}$.

3·2. *Estimating equations in the presence of misclassification*

When responses are subject to misclassification, the estimating functions in (2) and (3) with Y_{ij} replaced by the observed surrogate S_{ij} are no longer unbiased, and the resulting analysis usually yields inconsistent estimates of β and α (Yi & Reid, 2010). To conduct valid inference, one must correct the bias induced by misclassification. We propose modified estimating functions $U_{1i}^*(\theta)$ and $U_{2i}^*(\theta)$ using the observed data (S_i, X_i) so that their conditional expectations recover those in (2) and (3), i.e.,

$$E\{U_{1i}^*(\theta) \mid Y_i, X_i\} = U_{1i}(\theta), \quad E\{U_{2i}^*(\theta) \mid Y_i, X_i\} = U_{2i}(\theta). \tag{4}$$

Unbiasedness of $U_{si}^*(\theta)$ is immediate from that of $U_{si}(\theta)$ $(s = 1, 2)$. As a result, by the estimating function theory, under mild regularity conditions, solving

$$\sum_{i=1}^{n} \begin{pmatrix} U_{1i}^*(\theta) \\ U_{2i}^*(\theta) \end{pmatrix} = 0$$

gives a consistent estimator for θ.

Now we describe a way to construct $U_{si}^*(\theta)$ for $s = 1, 2$. Recognizing that response components appear in $U_{1i}(\theta)$ and $U_{2i}(\theta)$ merely through the linear term Y_{ij} and pairwise product $Z_{ijj'} = Y_{ij}Y_{ij'}$, we construct unbiased surrogates for Y_{ij} and $Z_{ijj'}$, namely

$$Y_{ij}^* = \frac{S_{ij} - 1 + \tau_{0ij}}{\tau_{0ij} + \tau_{1ij} - 1}, \quad Z_{ijj'}^* = \frac{a_0 + (S_{ij} - a_1)(S_{ij'} - a_2)}{a_3},$$

where

$$
\begin{aligned}
a_0 &= (1-a_1)\tau_{0ij'} + (1-a_2)\tau_{0ij} - \zeta_{ijj'}(0,0) - (1-a_1)(1-a_2),\\
a_1 &= \{\tau_{0ij} + \tau_{0ij'} + \tau_{1ij'} - 1 - \zeta_{ijj'}(0,1) - \zeta_{ijj'}(0,0)\}/(\tau_{1ij'} + \tau_{0ij'} - 1),\\
a_2 &= \{\tau_{0ij'} + \tau_{0ij} + \tau_{1ij} - 1 - \zeta_{ijj'}(1,0) - \zeta_{ijj'}(0,0)\}/(\tau_{1ij} + \tau_{0ij} - 1),\\
a_3 &= a_0 + a_1a_2 - a_1\tau_{1ij'} - a_2\tau_{1ij} + \zeta_{ijj'}(1,1).
\end{aligned}
$$

It is readily shown that $E(Y_{ij}^* \mid Y_i, X_i) = Y_{ij}$, and $E(Z_{ijj'}^* \mid Y_i, X_i) = Z_{ijj'}$ for $j \neq j'$. Let $Y_i^* = (Y_{i1}^*, \ldots, Y_{im_i}^*)^{\mathrm{T}}$, and $Z_i^* = (Z_{ijj'}^*, j < j')^{\mathrm{T}}$. Define

$$
\begin{pmatrix} U_{1i}^*(\theta) \\ U_{2i}^*(\theta) \end{pmatrix} = \begin{pmatrix} D_{1i}V_{1i}^{-1}\epsilon_{1i}^* \\ D_{2i}V_{2i}^{-1}\epsilon_{2i}^* \end{pmatrix}, \tag{5}
$$

where $\epsilon_{1i}^* = Y_i^* - \mu_i$ and $\epsilon_{2i}^* = Z_i^* - \xi_i$. It is immediate that $U_{1i}^*(\theta)$ and $U_{2i}^*(\theta)$ satisfy (4) because, given Y_i and X_i, the conditional expectation of Y_i^* and Z_i^* equals Y_i and Z_i, respectively.

We note that parameter η for the misclassification process comes into play in constructing Y_i^* and Z_i^*. Now we explicitly indicate this by writing $U_i^*(\theta, \eta) = \{U_{1i}^{*\mathrm{T}}(\theta, \eta), U_{2i}^{*\mathrm{T}}(\theta, \eta)\}^{\mathrm{T}}$. If η is known to be η_0, say, then under regularity conditions, solving the estimating equations

$$
\sum_{i=1}^{n} U_i^*(\theta, \eta_0) = 0
$$

leads to a consistent estimator, say $\hat{\theta}$, for θ. Under suitable regularity conditions, $n^{1/2}(\hat{\theta} - \theta)$ has an asymptotic normal distribution with mean 0 and covariance matrix $\Gamma_0^{*-1}\Sigma_0^*(\Gamma_0^{*-1})^{\mathrm{T}}$, where $\Gamma_0^*(\theta, \eta_0) = E\{\partial U_i^*(\theta, \eta_0)/\partial\theta^{\mathrm{T}}\}$, and $\Sigma_0^*(\theta, \eta_0) = E\{U_i^*(\theta, \eta_0)U_i^{*\mathrm{T}}(\theta, \eta_0)\}$. The proof is sketched in Appendix 1.

4. Inference method with a validation subsample available

In order to use (5) to perform inference about θ, it is critical that parameter η associated with misclassification is known. In practice, however, η is often unknown and must be estimated from an additional source of data. It is then important to accommodate induced variation in inferential procedures for θ. In this and next sections, we develop modified estimation algorithms to cover two practical situations, either a validation subsample or replicates of surrogates are available for estimation of η.

One may use the validation subsample to estimate the parameter η as well as to improve the efficiency of estimating θ for the mean and association parameters, as opposed to using surrogate observations for every subject. If the values of all misclassification indicators H_{ij}s were observed, estimates of η could be obtained as the solution to estimating equations (e.g., Lipsitz et al., 1991; Yi & Cook, 2002)

$$
\sum_{i=1}^{n} \begin{pmatrix} G_{1i}W_{1i}^{-1}e_{1i} \\ G_{2i}W_{2i}^{-1}e_{2i} \end{pmatrix} = 0,
$$

where $G_{1i} = \partial\tau_i^{\mathrm{T}}/\partial\gamma$, $G_{2i} = \partial\zeta_i^{\mathrm{T}}/\partial\nu$, $e_{1i} = H_i - \tau_i$, $e_{2i} = C_i - \zeta_i$, $W_{1i} = B_{\eta 1i}^{1/2}R_{\eta 1i}B_{\eta 1i}^{1/2}$, $B_{\eta 1i} = \mathrm{diag}[\tau_{i1}(y_{i1})\{1-\tau_{i1}(y_{i1})\}, \ldots, \tau_{im_i}(y_{im_i})\{1-\tau_{im_i}(y_{im_i})\}]$ and $R_{\eta 1i}$ is the correlation matrix of H_i. Analogous to V_{2i} in (3), matrix W_{2i} is often assumed to be the independence working matrix to avoid specification of higher order moments.

However, we do not observe the values of H_{ij}s unless subject j is in the validation subsample. Let $\delta_{ij}=1$ if the jth subject in cluster i belongs to the validation subsample and $\delta_{ij}=0$ otherwise. Here we assume that selection for a subject to be included in the validation subsample is noninformative. Let $\delta_i=(\delta_{i1},\ldots,\delta_{im_i})^{\mathrm{T}}$. Then estimating functions for η can be constructed from the measurements in the validation subsample. We now add a superscript δ to each vector and matrix to indicate the components corresponding to the validation subsample with $\delta_{ij}=1$. To be specific, let $Q_{1i}(\eta)=G_{1i}^{\delta}(W_{1i}^{\delta})^{-1}e_{1i}^{\delta}$, $Q_{2i}(\eta)=G_{2i}^{\delta}(W_{2i}^{\delta})^{-1}e_{2i}^{\delta}$, and $Q_i(\eta)=\{Q_{1i}^{\mathrm{T}}(\eta),\,Q_{2i}^{\mathrm{T}}(\eta)\}^{\mathrm{T}}$, then unbiased estimating equations for the η parameter are

$$\sum_{i=1}^{n} Q_i(\eta)=0. \tag{6}$$

In constructing valid estimating functions of θ, one can incorporate the available true response measurements in the validation subsample to improve efficiency, as opposed to using (5). To this end, we define $\tilde{Y}_{ij}=(1-\delta_{ij})Y_{ij}^{*}+\delta_{ij}Y_{ij}$, and $\tilde{Z}_{ijj'}=\{1-(1-\delta_{ij})(1-\delta_{ij'})\}\tilde{Y}_{ij}\tilde{Y}_{ij'}+(1-\delta_{ij})(1-\delta_{ij'})Z_{ijj'}^{*}$. Thus, $\tilde{Y}_{ij}=Y_{ij}$ if the jth subject in cluster i is in the validation subsample, $\tilde{Y}_{ij}=Y_{ij}^{*}$ otherwise; $\tilde{Z}_{ijj'}=\tilde{Y}_{ij}\tilde{Y}_{ij'}$ if either Y_{ij} or $Y_{ij'}$ or both are available, and $\tilde{Z}_{ijj'}=Z_{ijj'}^{*}$ otherwise. Denote $\tilde{Y}_i=(\tilde{Y}_{i1},\ldots,\tilde{Y}_{im_i})^{\mathrm{T}}$ and $\tilde{Z}_i=(\tilde{Z}_{ijj'},\,j<j')^{\mathrm{T}}$. Define $\tilde{U}_{1i}(\theta,\eta)=D_{1i}V_{1i}^{-1}\tilde{\epsilon}_{1i}$, and $\tilde{U}_{2i}(\theta,\eta)=D_{2i}V_{2i}^{-1}\tilde{\epsilon}_{2i}$, where $\tilde{\epsilon}_{1i}=\tilde{Y}_i-\mu_i$ and $\tilde{\epsilon}_{2i}=\tilde{Z}_i-\xi_i$. Let $\tilde{U}_i(\theta,\eta)=\{\tilde{U}_{1i}^{\mathrm{T}}(\theta,\eta),\tilde{U}_{2i}^{\mathrm{T}}(\theta,\eta)\}^{\mathrm{T}}$, then the augmented estimating equations for θ are

$$\sum_{i=1}^{n}\tilde{U}_i(\theta,\eta)=0. \tag{7}$$

Consistent estimators for η and θ can be obtained by jointly solving (6) and (7). Details are given in Appendix 2, where we also account for variation induced by the estimation of η.

5. JOINT ESTIMATION AND INFERENCE WITH REPLICATES

In this section we describe an inferential procedure for the case with replicates. Here we use notation slightly different from that in the previous sections. Let S_{ijr} be the rth replicate measure for Y_{ij}, $r=1,\ldots,d_{ij}$, where d_{ij} is the number of replicates for subject j in cluster i, $j=1,\ldots,m_i$, $i=1,\ldots,n$. Let $S_{ij}=(S_{ij1},\ldots,S_{ijd_{ij}})^{\mathrm{T}}$, and $H_{ijr}=\mathrm{I}(S_{ijr}=Y_{ij})$ be the misclassification indicator variable. For $(j,r)\neq(j',r')$, conditional independence between H_{ijr} and $H_{ij'r'}$ is assumed, given Y_i and X_i. For $r\neq r'$, H_{ijr} and $H_{ijr'}$ are assumed to have the same conditional distribution, given Y_i and X_i. Also it is assumed that $\mathrm{pr}(H_{ijr}=h_{ijr}\mid Y_i,X_i)=\mathrm{pr}(H_{ijr}=h_{ijr}\mid Y_{ij},X_i)$. Let $\tau_{1ijr}=\mathrm{pr}(H_{ijr}=1\mid Y_{ij}=1,X_i)$ and $\tau_{0ijr}=\mathrm{pr}(H_{ijr}=1\mid Y_{ij}=0,X_i)$. Suppose that τ_{1ijr} and τ_{0ijr} are modelled by (1).

Define $\mathcal{Y}_{ijr}^{*}=(S_{ijr}-1+\tau_{0ijr})/(\tau_{0ijr}+\tau_{1ijr}-1)$. Then the average version $\mathcal{Y}_{ij}^{*}=\sum_{r=1}^{d_{ij}}\mathcal{Y}_{ijr}^{*}/d_{ij}$ is unbiased for Y_{ij}, i.e., $E(\mathcal{Y}_{ij}^{*}\mid Y_i,X_i)=Y_{ij}$. Let $\mathcal{Y}_i^{*}=(\mathcal{Y}_{i1}^{*},\ldots,\mathcal{Y}_{im_i}^{*})^{\mathrm{T}}$, and $\mathcal{Z}_i^{*}=(\mathcal{Y}_{ij}^{*}\mathcal{Y}_{ij'}^{*},\,j<j')^{\mathrm{T}}$. Define $\mathcal{U}_{1i}(\theta,\gamma)=D_{1i}V_{1i}^{-1}\varepsilon_{1i}$, and $\mathcal{U}_{2i}(\theta,\gamma)=D_{2i}V_{2i}^{-1}\varepsilon_{2i}$, where $\varepsilon_{1i}=\mathcal{Y}_i^{*}-\mu_i$ and $\varepsilon_{2i}=\mathcal{Z}_i^{*}-\xi_i$ are residual vectors. It is readily seen that $E\{\mathcal{U}_{1i}(\theta,\gamma)\mid Y_i,X_i\}=U_{1i}(\theta,\gamma)$ and $E\{\mathcal{U}_{2i}(\theta,\gamma)\mid Y_i,X_i\}=U_{2i}(\theta,\gamma)$. If $\mathcal{U}_i(\theta,\gamma)=\{\mathcal{U}_{1i}^{\mathrm{T}}(\theta,\gamma),\mathcal{U}_{2i}^{\mathrm{T}}(\theta,\gamma)\}^{\mathrm{T}}$, then a consistent estimator of θ can be obtained by solving

$$\sum_{i=1}^{n}\mathcal{U}_i(\theta,\gamma)=0,$$

provided γ is given.

However, γ is unknown here, and must be estimated. In the case with replicates S_{ijr}, estimation of γ and θ typically interacts, and a joint estimation procedure is required to simultaneously estimate γ and θ. We generalize the discussion in White et al. (2001), who considered univariate logistic regression models with a misclassified binary covariate. Let $A_{ijk} = 1$ if $\sum_{r=1}^{d_{ij}} S_{ijr} = k$ and $A_{ijk} = 0$ otherwise, $k = 1, \ldots, d_{ij}$, $j = 1, \ldots, m_i$, $i = 1, \ldots, n$. Define $A_{ij} = (A_{ij1}, \ldots, A_{ijd_{ij}})^{\mathrm{T}}$, and $A_i = (A_{i1}^{\mathrm{T}}, \ldots, A_{im_i}^{\mathrm{T}})^{\mathrm{T}}$. Let $\pi_{ijk} = E(A_{ijk} \mid X_i)$ be the marginal mean of A_{ijk}. Let $\pi_{ij} = (\pi_{ij1}, \ldots, \pi_{ijd_{ij}})^{\mathrm{T}}$, and $\pi_i = (\pi_{i1}^{\mathrm{T}}, \ldots, \pi_{im_i}^{\mathrm{T}})^{\mathrm{T}}$.

Now we describe estimating functions for γ. For ease of exposition, we consider the case with $d_{ij} = 2$. The method can be easily extended to cases with $d_{ij} \geqslant 3$. Noting that

$$\begin{aligned}\mathrm{pr}(A_{ij1} = 1 \mid Y_i, X_i) &= \{(1 - \tau_{1ij1})\tau_{1ij2} + (1 - \tau_{1ij2})\tau_{1ij1}\}Y_{ij} \\ &\quad + \{(1 - \tau_{0ij1})\tau_{0ij2} + (1 - \tau_{0ij2})\tau_{0ij1}\}(1 - Y_{ij}), \\ \mathrm{pr}(A_{ij2} = 1 \mid Y_i, X_i) &= \tau_{1ij1}\tau_{1ij2}Y_{ij} + (1 - \tau_{0ij1})(1 - \tau_{0ij2})(1 - Y_{ij}),\end{aligned}$$

we write the marginal means of A_{ij1} and A_{ij2} as

$$\begin{aligned}\pi_{ij1} &= \{(1 - \tau_{1ij1})\tau_{1ij2} + (1 - \tau_{1ij2})\tau_{1ij1}\}\mu_{ij} \\ &\quad + \{(1 - \tau_{0ij1})\tau_{0ij2} + (1 - \tau_{0ij2})\tau_{0ij1}\}(1 - \mu_{ij}), \\ \pi_{ij2} &= \tau_{1ij1}\tau_{1ij2}\mu_{ij} + (1 - \tau_{0ij1})(1 - \tau_{0ij2})(1 - \mu_{ij}),\end{aligned}$$

respectively. Define $\mathcal{Q}_i(\theta, \gamma) = \mathcal{G}_i \mathcal{W}_i^{-1}(A_i - \pi_i)$, where $\mathcal{G}_i = \partial \pi_i^{\mathrm{T}} / \partial \gamma$, and $\mathcal{W}_i = \mathrm{cov}(A_i \mid X_i)$.

Let $\Psi_i^*(\theta, \gamma) = \{\mathcal{Q}_i^{\mathrm{T}}(\theta, \gamma), \mathcal{U}_i^{\mathrm{T}}(\theta, \gamma)\}^{\mathrm{T}}$. Now we solve

$$\sum_{i=1}^{n} \Psi_i^*(\theta, \gamma) = 0$$

for γ and θ using an iterative procedure. Details are given in Appendix 3, where we also establish asymptotic properties.

6. Numerical assessment of the proposed methods

6·1. *Design of simulation studies*

We conduct simulation studies to assess the performance of the proposed methods in contrast to the naive method which ignores misclassification. We consider a longitudinal study with $m_i = m = 3$ for $i = 1, \ldots, n$. The mean response model is given by

$$\mathrm{logit}(\mu_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3},$$

where $X_{ij1} = X_{i1}$ is 1 if the ith subject is randomized to the treatment group and 0 otherwise, and $X_{ij2} = \mathrm{I}(j = 2)$ and $X_{ij3} = \mathrm{I}(j = 3)$ describe temporal effects. An exchangeable structure,

$$\log \psi_{ijj'} = \alpha, \tag{8}$$

is considered for second-order association. The regression parameters are specified as $\exp(\beta_0) = 2$, $\exp(\beta_1) = 1/2$, $\exp(\beta_2) = 2/3$, and $\exp(\beta_3) = 1/3$, and the association parameter is specified

as $\alpha = \log(3)$. We generate binary response vectors from the joint distribution

$$\begin{aligned}&\mathrm{pr}(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, Y_{i3} = y_{i3})\\ &\quad = \prod_{j=1}^{3} \mu_{ij}^{y_{ij}} (1-\mu_{ij})^{1-y_{ij}} \left[1 + \sum_{1 \leqslant j < j' \leqslant 3} \rho_{ijj'} \frac{(y_{ij}-\mu_{ij})(y_{ij'}-\mu_{ij'})}{\{\mu_{ij}(1-\mu_{ij})\}^{1/2}\{\mu_{ij'}(1-\mu_{ij'})\}^{1/2}}\right],\end{aligned}$$

where $\rho_{ijj'}$ is the correlation coefficient between Y_{ij} and $Y_{ij'}$. This is a special case of the Bahadur representation (Bahadur, 1961), where only the marginal means and the second-order correlations are involved, and higher order correlations are constrained to be zero.

We consider both independent and correlated misclassification processes. For the independent case, we use a simple misclassification model that depends only on two parameters. The indicators H_{ij}s are generated under model

$$\mathrm{logit}(\tau_{ij}) = \begin{cases} \gamma_0 & (Y_{ij} = 0), \\ \gamma_1 & (Y_{ij} = 1). \end{cases} \tag{9}$$

Surrogate responses S_{ij} are then recorded as Y_{ij} if $H_{ij} = 1$ and $1 - Y_{ij}$ if $H_{ij} = 0$. Three settings for γ are considered:

(i) $\gamma_0 = \mathrm{logit}(0{\cdot}95)$ and $\gamma_1 = \mathrm{logit}(0{\cdot}95)$;
(ii) $\gamma_0 = \mathrm{logit}(0{\cdot}9)$ and $\gamma_1 = \mathrm{logit}(0{\cdot}9)$; and
(iii) $\gamma_0 = \mathrm{logit}(0{\cdot}8)$ and $\gamma_1 = \mathrm{logit}(0{\cdot}8)$,

which represent different levels of misclassification rates.

The performance of the proposed methods is assessed under three scenarios. For the first scenario where γ is known, each simulated sample contains $n = 200$ subjects. For the second scenario where γ is not known but an internal validation subsample is available, we take $n = 400$, and randomly select 30% of the subjects to be in the validation sample. For low misclassification rates as in setting (i), large sample size is usually necessary to obtain a valid estimate of γ. For the third scenario where γ is not known but replicates are available, the sample size is set to be $n = 200$ and two replicate surrogates are used for each Y_{ij}.

For cases where misclassifications within the same subject are correlated, the mean model is (9), and the association model is

$$\log\{\lambda_{ijj'}(y_{ij}, y_{ij'})\} = \nu_1 \mathrm{I}(y_{ij} = y_{ij'}) + \nu_2 \mathrm{I}(y_{ij} \neq y_{ij'}) \quad (1 \leqslant j < j' \leqslant 3; i = 1, \ldots, n).$$

We set $\nu_1 = \log(2)$ and $\nu_2 = \log(1{\cdot}5)$. Misclassification vector H_i is then generated from the joint model given by the Bahadur representation in the same manner as that for the response vector. Again, we consider two scenarios. In the first scenario we assume that η is given, and set the sample size to be $n = 200$. In the second scenario, η is treated as unknown and estimated. The sample size is set as $n = 400$ in this case, and 30% of subjects are randomly selected to be included in the validation subsample. For each specific combination of parameter values, we evaluate the performances of the approaches based on 2000 simulation replicates.

6·2. *Simulation results*

Table 1 shows the simulation results for the first and second scenarios for cases where the misclassification process is independent. The column under each approach represents the relative percentage bias, empirical variance, average of model-based variance, and coverage rate

Table 1. *Simulation results for the independent misclassification process*

	Naive method ($n = 200$)				Proposed method with known γ ($n = 200$)				Proposed method with unknown γ ($n = 400$)			
	Bias%†	EV	AMV	CP%	Bias%	EV	AMV	CP%	Bias%	EV	AMV	CP%
	(i) $\gamma_0 = \text{logit}(0{\cdot}95)$, $\gamma_1 = \text{logit}(0{\cdot}95)$											
β_0	−10·8	3·0	3·2	93·5	0·3	3·8	4·1	96·2	0·6	2·1	2·1	95·5
β_1	−9·6	3·8	4·0	93·8	2·0	4·9	5·1	95·2	0·4	2·4	2·5	95·5
β_2	−11·9	3·2	3·3	94·5	−1·2	4·0	4·2	95·7	−1·0	1·8	1·9	95·2
β_3	−10·6	3·5	3·6	90·3	0·8	4·6	4·7	95·6	0·7	2·1	2·2	95·8
α	−23·3	4·5	4·4	75·0	0·1	8·2	8·1	95·4	0·4	3·5	3·8	95·7
	(ii) $\gamma_0 = \text{logit}(0{\cdot}9)$, $\gamma_1 = \text{logit}(0 \cdot 9)$											
β_0	−21·1	2·9	3·1	86·6	1·0	5·0	5·2	95·7	0·2	2·7	2·7	95·6
β_1	−20·4	3·6	3·7	89·1	2·4	6·2	6·2	94·8	0·5	2·9	2·9	95·0
β_2	−21·9	3·5	3·5	92·3	−0·4	5·7	5·7	95·0	−1·6	2·3	2·5	96·5
β_3	−21·2	3·6	3·7	76·0	1·4	6·2	6·5	95·8	0·4	2·9	2·9	95·3
α	−41·8	4·1	4·0	36·6	0·9	14·0	13·8	95·4	1·0	6·1	6·1	95·5
	(iii) $\gamma_0 = \text{logit}(0{\cdot}8)$, $\gamma_1 = \text{logit}(0{\cdot}8)$											
β_0	−41·5	2·9	2·9	59·1	2·3	9·5	9·5	96·0	1·4	5·0	5·1	95·9
β_1	−41·3	3·2	3·2	63·6	3·5	10·4	10·4	95·3	1·2	4·5	4·5	95·8
β_2	−41·1	3·7	3·7	85·5	2·0	11·4	11·4	95·6	0·1	4·2	4·5	96·3
β_3	−41·6	3·9	3·9	36·5	3·1	13·1	13·3	95·8	1·9	5·3	5·6	96·1
α	−69·3	3·3	3·4	2·4	4·1	46·9	53·9	96·6	2·5	17·9	19·6	95·7

† Bias%, relative percentage bias, i.e., 100× (average estimate − true value)/true value; EV, empirical variance multiplied by 100; AMV, average of model-based variances multiplied by 100; CP%, coverage rate (%) of 95% confidence interval.

of 95% confidence intervals. We first look at the results under known γ. One can see that the naive analysis leads to downward biased estimates of response parameters even under a small proportion of misclassifications. Under setting (i) where the misclassification proportion is 5%, for example, both the mean parameters and the association parameter are attenuated by a non-ignorable amount. As the misclassification proportion increases, attenuation becomes more substantial. When the misclassification proportion is increased to 20% in setting (iii), coverage rates for the naive estimates of the mean parameters and association parameter are far below the nominal value 95%. In contrast, the proposed method performs reasonably well for all parameter configurations. For settings (i) and (ii) with small and moderate misclassification rates, the relative biases associated with the estimates of the mean parameters are fairly small. Relative biases increase slightly as the misclassification rate becomes higher. The coverage rate for α is slightly over the nominal 95%. The variance estimates of the estimators are larger than those of the naive estimators and increase as the misclassification rate increases, but they agree reasonably well with the empirical ones. For the case of estimated γ, similar patterns are observed.

Simulation results for the case with replicates are shown in Table 2. The relative biases and coverage rates for the naive estimators are similar to those in Table 1. The proposed method performs well. The relative biases of the estimates are small for the first two settings where misclassification rates are low and moderate, and the coverage rates are close to the nominal 95%. For setting (iii) with the highest misclassification rate, the relative biases in both the mean parameters and the association parameter are the largest. The coverage rate is slightly over the nominal 95% for the association parameter.

The results for correlated misclassifications are reported in Table 3. It is seen that ignoring misclassification leads to seriously biased estimates and considerably low coverage rates. The

Table 2. *Simulation results for the case with replicates where the misclassification process is independent*

	Naive method using the 1st replicates ($n = 200$)				Naive method using the 2st replicates ($n = 200$)				Proposed method ($n = 200$)			
	Bias%†	EV	AMV	CP%	Bias%	EV	AMV	CP%	Bias%	EV	AMV	CP%
	(i) $\gamma_0 = \text{logit}(0{\cdot}95)$, $\gamma_1 = \text{logit}(0{\cdot}95)$											
β_0	−9·9	3·2	3·2	92·6	−9·9	3·1	3·2	92·8	1·6	4·7	4·9	95·2
β_1	−10·0	4·0	4·0	92·9	−10·3	4·0	4·0	93·3	1·7	4·8	4·8	94·5
β_2	−10·5	3·4	3·3	94·1	−10·5	3·3	3·3	94·4	0·8	3·7	3·7	95·1
β_3	−10·6	3·6	3·6	90·7	−10·2	3·6	3·6	90·8	1·1	4·1	4·1	95·9
α	−22·4	4·3	4·4	76·5	−21·9	4·6	4·4	76·5	1·3	6·3	6·4	95·3
	(ii) $\gamma_0 = \text{logit}(0{\cdot}9)$, $\gamma_1 = \text{logit}(0{\cdot}9)$											
β_0	−20·7	3·2	3·1	86·0	−21·1	3·0	3·1	86·6	2·6	7·7	7·8	94·9
β_1	−21·0	3·8	3·7	87·1	−20·9	3·8	3·7	88·0	2·0	5·4	5·3	94·9
β_2	−21·0	3·7	3·5	91·2	−21·1	3·4	3·5	92·5	0·7	4·5	4·5	95·2
β_3	−21·7	4·0	3·7	74·3	−21·5	3·6	3·7	76·2	0·9	5·0	4·9	95·0
α	−41·5	3·9	4·0	36·6	−41·3	4·3	4·0	37·5	1·3	8·7	8·8	94·8
	(iii) $\gamma_0 = \text{logit}(0{\cdot}8)$, $\gamma_1 = \text{logit}(0{\cdot}8)$											
β_0	−40·9	2·9	2·9	60·8	−41·8	2·8	2·9	60·0	5·1	22·2	23·5	95·9
β_1	−41·5	3·3	3·2	63·3	−42·6	3·3	3·2	61·8	3·7	7·7	7·6	94·7
β_2	−41·7	3·7	3·7	85·4	−41·1	3·7	3·7	85·9	2·8	7·5	7·6	95·5
β_3	−42·3	4·1	3·9	35·1	−41·6	3·8	3·9	36·5	3·5	8·8	8·7	95·5
α	−69·3	3·7	3·4	2·6	−68·6	3·5	3·4	2·8	6·8	24·2	24·8	96·6

† Bias%, relative percentage bias, i.e., 100× (average estimate − true value)/true value; EV, empirical variance multiplied by 100; AMV, average of model-based variances multiplied by 100; CP%, coverage rate (%) of 95% confidence interval.

Table 3. *Simulation results for the correlated misclassification process*

	Naive method ($n = 200$)				Proposed method with known η ($n = 200$)				Proposed method with unknown η ($n = 400$)			
	Bias%†	EV	AMV	CP%	Bias%	EV	AMV	CP%	Bias%	EV	AMV	CP%
	(i) $\gamma_0 = \text{logit}(0 \cdot 95)$, $\gamma_1 = \text{logit}(0{\cdot}95)$											
β_0	−10·4	3·0	3·2	93·2	0·8	3·5	3·7	95·4	0·1	1·9	1·9	95·2
β_1	−10·9	3·8	4·0	94·1	0·4	4·5	4·7	95·7	0·1	2·3	2·4	95·2
β_2	−10·0	3·2	3·3	94·4	1·1	3·6	3·7	95·1	0·2	1·8	1·8	94·5
β_3	−10·2	3·6	3·6	90·5	1·0	4·2	4·1	94·8	0·3	2·0	2·1	95·7
α	−22·2	4·3	4·4	76·6	−1·0	9·9	9·8	95·2	0·6	3·3	3·4	95·4
	(ii) $\gamma_0 = \text{logit}(0{\cdot}9)$, $\gamma_1 = \text{logit}(0{\cdot}9)$											
β_0	−21·5	1·5	1·5	77·3	0·2	2·1	2·1	95·4	0·2	2·2	2·3	95·4
β_1	−21·7	1·9	1·9	80·1	0·4	2·7	2·7	94·4	0·4	2·7	2·7	94·2
β_2	−21·9	1·7	1·7	89·4	−0·8	2·1	2·2	95·2	−0·8	2·1	2·2	95·3
β_3	−21·6	1·8	1·8	57·5	0·4	2·4	2·4	95·0	0·4	2·4	2·5	95·4
α	−36·5	2·1	2·0	20·4	0·8	8·9	8·5	94·9	1·1	5·4	5·3	94·2
	(iii) $\gamma_0 = \text{logit}(0{\cdot}8)$, $\gamma_1 = \text{logit}(0{\cdot}8)$											
β_0	−42·2	1·5	1·5	33·4	0·1	3·1	3·2	95·4	−0·1	3·5	3·6	95·9
β_1	−42·5	1·7	1·7	38·4	0·6	3·7	3·8	94·7	0·6	3·8	3·8	95·0
β_2	−42·6	1·8	1·8	74·2	−1·3	3·5	3·4	95·0	−1·2	3·5	3·5	95·4
β_3	−42·3	1·9	1·9	7·6	0·5	4·0	4·0	95·0	0·6	4·2	4·2	95·0
α	−57·0	1·7	1·8	0·3	0·7	24·6	25·1	96·4	0·8	16·9	17·1	95·9

†Bias%, relative percentage bias, i.e., 100× (average estimate − true value)/true value; EV, empirical variance multiplied by 100; AMV, average of model-based variances multiplied by 100; CP%, coverage rate of 95% confidence interval in percent.

proposed approach yields quite satisfactory estimates of the mean and association parameters, regardless whether parameter η is taken as known or estimated.

In summary, the simulation studies demonstrate that the proposed method works well in various situations, and it produces reliable point estimates as well as standard errors for both mean and association parameters governing the response process. Finally, we comment on a numerical issue related to estimation of association parameters ν for the misclassification process. If the size of a validation subsample is small, then estimation of ν could be unstable, which in turn influences estimation of the response parameters. In this case, we may ignore possible correlation between misclassification indicators but just model the marginal misclassification probabilities. Our numerical experience shows that this approach can help overcome instability of estimation of the response parameters.

7. Application

We apply the proposed method to a dataset arising from the Canadian Community Health Survey cycle 3·1 conducted in 2005. This is a large-scale on-going survey targeting individuals aged 12 and older in the Canadian population. The design of the survey is fairly complex, with three sampling frames being used to sample households: an area frame, a list frame of telephone numbers, and a random digit dialing sampling frame. For each sampled household, an individual aged 12 and older was randomly chosen for interview.

The objective of our study is to explore the relationship between obesity and some risk factors. We consider a sample of 2699 respondents aged 18 and older in the Toronto health region. These respondents were from 435 clusters based on postal codes with size varying from 2 to 15. Among them, 150 were included by randomization as a validation subsample for which body mass index was directly measured, and the resultant obesity status was regarded as the true response value for each subject in this subsample (Shields et al., 2008). For other individuals, the obesity status was determined by the self-reported information, and therefore was subject to error. Covariates include age, sex, and physical activity index. There are three levels of physical activity index: active, moderate, which is treated as a reference category, and inactive. Let Y_{ij} denote the binary obesity status for subject j in cluster i. We assume that Y_{ij} follows the logistic model

$$\text{logit}\,\mu_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_4 X_{ij4},$$

where X_{ij1} is the subject's age, X_{ij2} is 1 if the subject is male and 0 otherwise, X_{ij3} is 1 if physical activity index is active and 0 otherwise, and X_{ij4} is 1 if the physical activity index is inactive and 0 otherwise. The association between Y_{ij} and $Y_{ij'}$, measured by odds ratio $\psi_{ijj'}$, is modelled by (8). Because the surrogate responses are obtained from self-reporting, misclassifications in obesity are typically independent for different individuals and clusters.

We conducted analyses under two different assumptions for the misclassification process, the first assuming that misclassification is independant of covariates, and the second assuming age-dependent misclassification. Table 4 shows the results for the first analysis. We also report results from a naive analysis ignoring misclassification. The estimated coefficient of age is 0·016, indicating that older subjects have higher probability of developing obesity. The probability of developing obesity is not significantly different between males and females at the 5% level. There is some evidence that active subjects have a smaller chance of developing obesity compared to moderately active subjects. In contrast, the subjects in the inactive group are more likely to develop obesity compared the those in the other groups. The association parameter α is estimated to be 0·106, which corresponds to an odds ratio of 1·11 between obesities of two subjects

Table 4. *Analysis with misclassification independent of covariates*

		Naive method			Proposed method		
		Est.	SE	p-value	Est.	SE	p-value
Response models							
Intercept		−2·798	0·225	<0·001	−2·652	0·372	<0·001
Age		0·014	0·003	<0·001	0·016	0·005	<0·001
Sex	male	0·006	0·124	0·958	0·003	0·152	0·982
Activity	active	−0·421	0·191	0·027	−0·550	0·265	0·038
	inactive	0·345	0·153	0·025	0·427	0·189	0·024
Association: (α)		0·073	0·114	0·521	0·106	0·170	0·532
Misclassification models							
$\text{pr}(S = 0 \mid Y = 0)$					0·984	0·712	0·076
$\text{pr}(S = 1 \mid Y = 1)$					0·667	0·408	<0·001

Est., estimate; SE, standard error.

Table 5. *Analysis with age-dependent misclassification*

		Naive method			Proposed method		
		Est.	SE	p-value	Est.	SE	p-value
Response models							
Intercept		−2·798	0·225	<0·001	−4·356	1·113	<0·001
Age		0·014	0·003	<0·001	0·048	0·022	0·028
Sex	male	0·006	0·124	0·958	−0·057	0·197	0·771
Activity	active	−0·421	0·191	0·027	−0·211	0·397	0·596
	inactive	0·345	0·153	0·025	0·731	0·351	0·038
Association: (α)		0·073	0·114	0·521	0·124	0·228	0·585
Misclassification models							
$\text{pr}(S = 0 \mid Y = 0)$							
	Intercept				3·885	2·063	0·060
	Age				0·005	0·046	0·912
$\text{pr}(S = 1 \mid Y = 1)$							
	Intercept				5·904	2·324	0·011
	Age				−0·094	0·039	0·017

Est., estimate; SE, standard error.

in the same cluster. However, there is no strong evidence for this association. Compared with the proposed method, the naive approach generally produced attenuated estimates of the regression parameters. Table 5 shows the results for the second analysis. There is no evidence that age is associated with misclassification from non-obesity to obesity. However, age is significantly associated with misclassification from obesity to non-obesity at the 5% level. Older people who are obese tend to underestimate their body mass index, leading to false self-reported non-obesity. The probability of developing obesity is not significantly different between active subjects and moderately active subjects, while conclusions for the other covariates remain the same. Comparing Table 5 with Table 4 we observe some inflation of standard errors associated with the estimated regression parameters in the response model, perhaps because the validation subsample is relatively small. This phenomenon has been a long concern in correction for misclassification in binary responses, e.g., Luan et al. (2005). When the validation subsample contains a very small number of misclassifications, it may be preferred not to fit a complex misclassification model involving covariates, because parameter estimates are usually unstable under such scenarios.

8. Discussion

Misclassification arises commonly with binary or categorial data. Ignoring it often results in biased inference. In this paper we propose a method to correct for bias induced by misclassified binary responses with a complex association structure. We focus on modelling both marginal mean and association structures for the response process, and develop a marginal analysis method based on unbiased estimating functions that are expressed in terms of surrogate responses along with other observed measurements. The proposed method in terms of surrogate responses along with other observed measurements. The proposed method is motivated by the unique feature of the generalized estimating functions U_{1i} and U_{2i} under the situation that no misclassification is present. Recognizing that response components appear in U_{1i} and U_{2i} solely via linear and cross-product terms, Y_{ij} and $Z_{ijj'}$, we construct unbiased surrogates Y^*_{ij} and $Z^*_{ijj'}$ and use them to respectively replace Y_{ij} and $Z_{ijj'}$ in U_{1i} and U_{2i}. Yi's and Wu's replacement differs from the naive method that directly substitutes the Y_{ij} with S_{ij}. This approach can not only make the resulting estimating functions computable but also preserve their unbiasedness. The proposed method is attractive in that it is conceptually simple and easy to implement. Furthermore, it makes the best use of the model setup without requiring additional distributional assumptions.

Since misclassification parameters are often unknown, additional information such as validation data or replicated measures is often needed to obtain estimates of these parameters. A validation subsample may be available in two-stage designs for which cost- and time-efficiency would be a primary concern. In those studies, the (S, X) are measured for al subjects at the first stage of the study, and in the second stage, Y is also measured for a subset of the study participants. In other situations such as the survey context, validation data can be collected by assigning more experienced interviewers to a subset of randomly selected individuals, while other individuals' measurements are obtained based on self-reported questionnaires. In circumstances where no validation data nor replicates are available, one may conduct sensitivity analyses to evaluate the impact of misclassification on inference about the response parameters. The method discussed in § 3·2 can be applied for this purpose.

Acknowledgement

The authors are grateful to Statistics Canada/The Research Data Centre Network for the permission to use the data from the Canadian Community Health Survey. The authors thank the editor, associate editor, and the two referees for their helpful comments. The authors also thank Drs. Richard Cook and Mary Thompson for providing helpful comments on an earlier draft of this manuscript. Yi's and Wu's research was supported by grants from the Natural Sciences and Engineering Research Council of Canada. Chen was partially by an internship fund from Mathematics of Information Technology and Complex Systems.

Appendix 1

Consistency and asymptotic distribution of $\hat{\theta}$

Because $U_i^*(\theta, \eta_0)$ satisfies $E\{U_i^*(\theta, \eta_0) \mid X_i\} = 0$ at the true value of θ, by Theorem 3.4 of Newey & McFadden (1993) and under the regularity conditions specified there, we have that with probability approaching 1, there is a unique solution, denoted by $\hat{\theta}$, to $\sum_{i=1}^n U_i^*(\theta, \eta_0) = 0$ that satisfies

$$0 = n^{-1/2} \sum_{i=1}^{n} U_i^*(\theta, \eta_0) + n^{-1} \sum_{i=1}^{n} \partial U_i^*(\theta, \eta_0)/\partial \theta^{\mathrm{T}} n^{1/2}(\hat{\theta} - \theta) + o_p(1).$$

This is equivalent to

$$n^{1/2}(\hat{\theta}-\theta) = -[E\{\partial U_i^*(\theta,\eta_0)/\partial\theta^{\mathrm{T}}\}]^{-1} n^{-1/2}\sum_{i=1}^{n} U_i^*(\theta,\eta_0) + o_p(1) \tag{A1}$$

as under regularity conditions, $E\{\partial U_i^*(\theta,\eta_0)/\partial\theta^{\mathrm{T}}\}$ exists and is invertible and $\mathrm{var}\{U_i^*(\theta,\eta_0)\}$ is finite and positive definite. The law of large numbers leads to that $n^{-1}\sum_{i=1}^{n} U_i^*(\theta,\eta_0)$ converges to $E\{U_i^*(\theta,\eta_0)\} = 0$ in probability as n goes to infinity, and the consistency of $\hat{\theta}$ is immediate by Slutzky's lemma. By applying the central limit theorem to (A1), the asymptotic distribution of $n^{1/2}(\hat{\theta}-\theta)$ can be established.

APPENDIX 2

Consistency and asymptotic distribution of $\hat{\theta}_v$

Define $\Psi_i(\theta,\eta) = \{Q_i^{\mathrm{T}}(\eta), \tilde{U}_i^{\mathrm{T}}(\theta,\eta)\}^{\mathrm{T}}$. Then solving

$$\sum_{i=1}^{n}\Psi_i(\theta,\eta) = 0 \tag{A2}$$

using the Fisher scoring algorithm yields consistent estimators for η and θ. To be specific, let

$$J_i(\eta) = E\left\{\frac{\partial Q_i^{\mathrm{T}}(\eta)}{\partial\eta}\right\} = \begin{pmatrix} J_{1i}(\eta) & 0 \\ J_{21i}(\eta) & J_{2i}(\eta)\end{pmatrix}, \quad M_i(\theta) = E\left\{\frac{\partial\tilde{U}_i^{\mathrm{T}}(\theta,\eta)}{\partial\theta}\right\} = \begin{pmatrix} M_{1i}(\theta) & 0 \\ M_{21i}(\theta) & M_{2i}(\theta)\end{pmatrix},$$

where $J_{1i}(\eta) = -G_{1i}^{\delta}(W_{1i}^{\delta})^{-1}\left(G_{1i}^{\delta}\right)^{\mathrm{T}}$, $J_{21i}(\eta) = -G_{2i}^{\delta}(W_{2i}^{\delta})^{-1}(\partial\zeta_i^{\delta}/\partial\gamma^{\mathrm{T}})$, $J_{2i}(\eta) = -G_{2i}^{\delta}(W_{2i}^{\delta})^{-1}(G_{2i}^{\delta})^{\mathrm{T}}$, $M_{1i}(\theta) = -D_{1i}V_{1i}^{-1}D_{1i}^{\mathrm{T}}$, $M_{21i}(\theta) = -D_{2i}V_{2i}^{-1}(\partial\xi_i/\partial\beta^{\mathrm{T}})$, and $M_{2i}(\theta) = -D_{2i}V_{2i}^{-1}D_{2i}^{\mathrm{T}}$. Estimates of η and θ can be obtained via an iterative equation

$$\begin{pmatrix}\eta^{(t+1)}\\ \theta^{(t+1)}\end{pmatrix} = \begin{pmatrix}\eta^{(t)}\\ \theta^{(t)}\end{pmatrix} + \begin{pmatrix} -\sum_{i=1}^{n} J_i(\eta^{(t)}) & 0 \\ -\sum_{i=1}^{n} E\{\partial\tilde{U}_i(\theta,\eta)/\partial\eta^{\mathrm{T}}\}\,|_{(\theta^{(t)},\eta^{(t)})} & -\sum_{i=1}^{n} M_i(\theta^{(t)})\end{pmatrix}^{-1} \times\sum_{i=1}^{n}\Psi_i(\theta^{(t)},\eta^{(t)}) \quad (t=0,1,\ldots),$$

until convergence. Let $(\hat{\eta}_v,\hat{\theta}_v)$ denote the resulting limit.

Let $\tilde{\Omega}_i(\theta,\eta) = \tilde{U}_i(\theta,\eta) - E\{\partial\tilde{U}_i(\theta,\eta)/\partial\eta^{\mathrm{T}}\}[E\{\partial Q_i(\eta)/\partial\eta^{\mathrm{T}}\}]^{-1}Q_i(\eta)$, and $\tilde{\Gamma}(\theta,\eta) = E\{\partial\tilde{U}_i(\theta,\eta)/\partial\theta^{\mathrm{T}}\}$. Following arguments similar to those in Chen et al. (2010), because $E\{\Psi_i(\theta,\eta)\} = 0$ when the response and misclassification models are correctly specified, under standard regularity conditions there is a unique solution, $(\hat{\eta}_v,\hat{\theta}_v)$ to (A2), with probability approaching one. By the first-order Taylor series approximation, we have

$$n^{1/2}\begin{pmatrix}\hat{\eta}_v-\eta\\ \hat{\theta}_v-\theta\end{pmatrix} = -\begin{pmatrix} E\{\partial Q_i(\eta)/\partial\eta^{\mathrm{T}}\} & 0 \\ E\{\partial\tilde{U}_i(\theta,\eta)/\partial\eta^{\mathrm{T}}\} & E\{\partial\tilde{U}_i(\theta,\eta)/\partial\theta^{\mathrm{T}}\}\end{pmatrix}^{-1} \times n^{-1/2}\sum_{i=1}^{n}\Psi_i(\theta,\eta) + o_p(1).$$

It follows that $n^{1/2}(\hat{\theta}_v-\theta)$ equals

$$-n^{-1/2}[E\{\partial\tilde{U}_i(\theta,\eta)/\partial\theta^{\mathrm{T}}\}]^{-1}\left\{\sum_{i=1}^{n}\tilde{U}_i(\theta,\eta) - E\{\partial\tilde{U}_i(\theta,\eta)/\partial\eta^{\mathrm{T}}\} \times[E\{\partial Q_i(\eta)/\partial\eta^{\mathrm{T}}\}]^{-1}\sum_{i=1}^{n}Q_i(\eta)\right\} + o_p(1) = -n^{-1/2}\tilde{\Gamma}^{-1}(\theta,\eta)\sum_{i=1}^{n}\tilde{\Omega}_i(\theta,\eta) + o_p(1).$$

Then applying the central limit theorem, one can show that $n^{1/2}(\hat{\theta}_v - \theta)$ is asymptotically normally distributed with mean 0 and asymptotic covariance matrix given by $\tilde{\Gamma}^{-1}\tilde{\Sigma}(\tilde{\Gamma}^{-1})^{\mathrm{T}}$, where $\tilde{\Sigma} = E\{\tilde{\Omega}_i(\theta,\eta)\tilde{\Omega}_i^{\mathrm{T}}(\theta,\eta)\}$.

Define $\tilde{\Lambda}_i(\theta,\eta) = [\{D_{1i}V_{1i}^{-1}(\partial\tilde{Y}_i/\partial\eta^{\mathrm{T}})\}^{\mathrm{T}}, \{D_{2i}V_{2i}^{-1}(\partial\tilde{Z}_i/\partial\eta^{\mathrm{T}})\}^{\mathrm{T}}]^{\mathrm{T}}$. As n goes to infinity, $E\{\partial\tilde{U}_i(\theta,\eta)/\partial\eta^{\mathrm{T}}\}$ and $E\{\partial Q_i(\eta)/\partial\eta^{\mathrm{T}}\}$ can be consistently estimated by $\tilde{\Lambda}(\hat{\theta}_v,\hat{\eta}_v) = n^{-1}\sum_{i=1}^n \tilde{\Lambda}_i(\hat{\theta}_v,\hat{\eta}_v)$ and $J(\hat{\eta}_v) = n^{-1}\sum_{i=1}^n J_i(\hat{\eta}_v)$, respectively. The matrices $\tilde{\Sigma}$ and $\tilde{\Gamma}$ can be consistently estimated by $\hat{\Sigma} = n^{-1}\sum_{i=1}^n \tilde{\Omega}_i(\hat{\theta}_v,\hat{\eta}_v)\tilde{\Omega}_i^{\mathrm{T}}(\hat{\theta}_v,\hat{\eta}_v)$ and $\hat{\Gamma} = n^{-1}\sum_{i=1}^n M_i(\hat{\theta}_v)$, respectively, where $\tilde{\Omega}_i(\hat{\theta}_v,\hat{\eta}_v) = \tilde{U}_i(\hat{\theta}_v,\hat{\eta}_v) - \tilde{\Lambda}(\hat{\theta}_v,\hat{\eta}_v)J^{-1}(\hat{\eta}_v)Q_i(\hat{\eta}_v)$. Therefore, the empirical version $n^{-1}\hat{\Gamma}^{-1}\hat{\Sigma}(\hat{\Gamma}^{-1})^{\mathrm{T}}$ is a consistent estimator for the asymptotic covariance matrix of $\hat{\theta}$.

Appendix 3

Consistency and asymptotic distribution of $\hat{\theta}_r$

Let $\mathcal{J}_i(\theta,\gamma) = -\mathcal{G}_i\mathcal{W}_i^{-1}\mathcal{G}_i^{\mathrm{T}}$, $\Delta_i(\theta,\gamma) = -\mathcal{G}_i\mathcal{W}_i^{-1}(\partial\pi_i/\partial\theta^{\mathrm{T}})$, and $\Lambda_i^*(\theta,\eta) = [\{D_{1i}V_{1i}^{-1}(\partial\mathcal{Y}_i^*/\partial\eta^{\mathrm{T}})\}^{\mathrm{T}}, \{D_{2i}V_{2i}^{-1}(\partial\mathcal{Z}_i^*/\partial\eta^{\mathrm{T}})\}^{\mathrm{T}}]^{\mathrm{T}}$. Given initial estimates $\theta^{(0)}$ and $\gamma^{(0)}$, we update the estimates via

$$\begin{pmatrix}\gamma^{(t+1)}\\ \theta^{(t+1)}\end{pmatrix} = \begin{pmatrix}\gamma^{(t)}\\ \theta^{(t)}\end{pmatrix} - \begin{pmatrix}\sum_{i=1}^n \mathcal{J}_{1i}(\theta^{(t)},\gamma^{(t)}) & \sum_{i=1}^n \Delta_i(\theta^{(t)},\gamma^{(t)})\\ \sum_{i=1}^n \Lambda_i(\theta^{(t)},\gamma^{(t)}) & \sum_{i=1}^n M_i(\theta^{(t)})\end{pmatrix}^{-1}\sum_{i=1}^n \Psi_i^*(\theta^{(t)},\gamma^{(t)}), \quad t = 0, 1, \ldots,$$

until convergence. Let $(\hat{\gamma}_r, \hat{\theta}_r)$ denote the resulting limit.

Note that $E\{\Psi_i^*(\theta,\gamma)\} = 0$ when both the response and misclassification models are correctly specified, since $E(A_i) = \pi_i$. Therefore $(\hat{\gamma}_r, \hat{\theta}_r)$ are consistent estimators for (γ,θ), and the asymptotic distribution of $\hat{\theta}_r$ can be established in a similar manner to that in § 4·2. However, there is an important difference arising from the interplay of θ and γ in both $\mathcal{U}_i(\theta,\gamma)$ and $\mathcal{Q}_i(\theta,\gamma)$. Specifically, applying the Taylor series expansion, we obtain

$$n^{1/2}\begin{pmatrix}\hat{\gamma}_r - \gamma\\ \hat{\theta}_r - \theta\end{pmatrix} = -\begin{pmatrix}E\{\partial\mathcal{Q}_i(\theta,\gamma)/\partial\gamma^{\mathrm{T}}\} & E\{\partial\mathcal{Q}_i(\theta,\gamma)/\partial\theta^{\mathrm{T}}\}\\ E\{\partial\mathcal{U}_i(\theta,\gamma)/\partial\gamma^{\mathrm{T}}\} & E\{\partial\mathcal{U}_i(\theta,\gamma)/\partial\theta^{\mathrm{T}}\}\end{pmatrix}^{-1} n^{-1/2}\sum_{i=1}^n \Psi_i^*(\theta,\gamma) + o_p(1).$$

It follows that $n^{1/2}(\hat{\theta}_r - \theta)$ equals

$$\begin{aligned}&n^{-1/2}(E\{\partial\mathcal{U}_i(\theta,\gamma)/\partial\theta^{\mathrm{T}}\} - E\{\partial\mathcal{U}_i(\theta,\gamma)/\partial\gamma^{\mathrm{T}}\}[E\{\partial\mathcal{Q}_i(\theta,\gamma)/\partial\gamma^{\mathrm{T}}\}]^{-1}E\{\partial\mathcal{Q}_i(\theta,\gamma)/\partial\theta^{\mathrm{T}}\})^{-1}\\ &\quad\times\left(\sum_{i=1}^n \mathcal{U}_i(\theta,\gamma) - E\{\partial\mathcal{U}_i(\theta,\gamma)/\partial\gamma^{\mathrm{T}}\}[E\{\partial\mathcal{Q}_i(\theta,\gamma)/\partial\gamma^{\mathrm{T}}\}]^{-1}\sum_{i=1}^n \mathcal{Q}_i(\theta,\gamma)\right) + o_p(1)\\ &\qquad = n^{-1/2}\Gamma^{*-1}(\theta,\gamma)\sum_{i=1}^n \Omega_i^*(\theta,\gamma) + o_p(1),\end{aligned}$$

where $\Omega_i^*(\theta,\gamma) = \mathcal{U}_i(\theta,\gamma) - E\{\partial\mathcal{U}_i(\theta,\gamma)/\partial\gamma^{\mathrm{T}}\}[E\{\partial\mathcal{Q}_i(\theta,\gamma)/\partial\gamma^{\mathrm{T}}\}]^{-1}\mathcal{Q}_i(\theta,\gamma)$, and $\Gamma^*(\theta,\gamma) = E\{\partial\mathcal{U}_i(\theta,\gamma)/\partial\theta^{\mathrm{T}}\} - E\{\partial\mathcal{U}_i(\theta,\gamma)/\partial\gamma^{\mathrm{T}}\}[E\{\partial\mathcal{Q}_i(\theta,\gamma)/\partial\gamma^{\mathrm{T}}\}]^{-1}E\{\partial\mathcal{Q}_i(\theta,\gamma)/\partial\theta^{\mathrm{T}}\}$. Thus, the Central Limit Theorem yields that $n^{1/2}(\hat{\theta}_r - \theta)$ is asymptotically normally distributed with mean 0 and covariance matrix $\Gamma^{*-1}\Sigma^*(\Gamma^{*-1})^{\mathrm{T}}$, where $\Sigma^* = E\{\Omega_i^*(\theta,\gamma)\Omega_i^{*\mathrm{T}}(\theta,\gamma)\}$.

As n goes to infinity, Γ^* and Σ^* can be consistently estimated by their empirical counterparts $\hat{\Gamma}^* = n^{-1}\sum_{i=1}^n[M_i(\hat{\theta}_r) - \Lambda_i^*(\hat{\theta}_r,\hat{\gamma}_r)\{\mathcal{J}_i(\hat{\theta}_r,\hat{\gamma}_r)\}^{-1}\Delta_i^*(\hat{\theta}_r,\hat{\gamma}_r)]$ and $\hat{\Sigma}^* = n^{-1}\sum_{i=1}^n \Omega_i^*(\hat{\theta}_r,\hat{\gamma}_r)\Omega_i^*(\hat{\theta}_r,\hat{\gamma}_r)^{\mathrm{T}}$, respectively. A consistent estimator for the asymptotic covariance matrix of $\hat{\theta}_r$ is given by $\hat{\Gamma}^{*-1}\hat{\Sigma}^*(\hat{\Gamma}^{*-1})^{\mathrm{T}}$.

References

Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. In *Studies in Item Analysis and Prediction*, Ed. H. Solomon, pp. 158–68, Stanford Mathematical Studies in the Social Sciences VI. Stanford, CA: Stanford University Press.

CAREY, V., ZEGER, S. L. & DIGGLE, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80**, 517–26.

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. & CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models*, 2nd ed. London: Chapman and Hall.

CHEN, B., YI, G. Y. & COOK, R. J. (2010). Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. *J. Am. Statist. Assoc.* **105**, 336–53.

COOK, R. J., NG, E. T. M. & MEADE, M. O. (2000). Estimation of operating characteristics for dependent diagnostic tests based on latent Markov models. *Biometrics* **56**, 1109–17.

LIANG, K.-Y. & ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

LIPSITZ, S. R., LAIRD, N. M. & HARRINGTON, D. P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika* **78**, 153–60.

LUAN, X., PAN, W., GERBERICH, S. G. & CARLIN, B. P. (2005). Does it always help to adjust for misclassification of a binary outcome in logistic regression? *Statist. Med.* **24**, 2221–34.

MILLER, M. E., DAVIS, C. S. & LANDIS, J. R. (1993). The analysis of longitudinal polytomous data: generalized estimating equations and connections with weighted least squares. *Biometrics* **49**, 1033–44.

MOLENBERGHS, G. & LESAFFRE, E. (1999). Marginal modelling of multivariate categorical data. *Statist. Med.* **18**, 2237–55.

NEUHAUS, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika* **86**, 843–55.

NEUHAUS, J. M. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics* **58**, 675–83.

NEWEY, W. K. & MCFADDEN, D. (1993). Estimation in large samples. In *Handbook of Econometrics*, Vol. 4. Ed. D. McFadden & R. Engler. Amsterdam: Holland.

PAULINO, C. D., SILVA, G. & ACHCAR, J. A. (2005). Bayesian analysis of correlated misclassified binary data. *Comp. Statist. Data Anal.* **49**, 1120–31.

PAULINO, C. D., SOARES, P. & NEUHAUS, J. (2003). Binomial regression with misclassification. *Biometrics* **59**, 670–75.

PEPE, M. & ANDERSON, G. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Commun. Statist.* **23**, 939–51.

PRENTICE, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–48.

ROSYCHUK, R. J. & THOMPSON, M. E. (2001). A semi-Markov model for binary longitudinal responses subject to misclassification. *Can. J. Statist.* **29**, 395–404.

ROSYCHUK, R. J. & THOMPSON, M. E. (2003). Bias correction of two-state latent Markov process parameter estimates under misclassification. *Statist. Med.* **22**, 2035–55.

SHIELDS, M., CONNOR GORBER, S. & TREMBLAY, M. S. (2008). Estimates of obesity based on self-report versus direct measures. *Health Reports* (Statistics Canada, Catalogue 82-003) **19**, 61–76.

WHITE, I., FROST, C. & TOKUNAGA, S. (2001). Correcting for measurement error in binary and continuous variables using replicates. *Statist. Med.* **20**, 3441–57.

YI, G. Y. & COOK, R. J. (2002). Marginal methods for incomplete longitudinal data arising in clusters. *J. Am. Statist. Assoc.* **97**, 1071–80.

YI, G. Y. & REID, N. (2010). A note on mis-specified estimating functions. *Statist. Sinica* **20**, 1749–69.

[*Received April* 2010. *Revised April* 2011]