Statistics and Its Interface Volume 12 (2019) 309–320

Analysis of panel data with misclassified covariates

Grace Yi[*], Wenqing He, and Feng He

Markov models are commonly used to describe the disease progression, and the likelihood method is usually used to perform inference for such models. However, in the presence of measurement error in the variables, standard inference procedures are no longer valid. In this article, we analytically show that the model is not even identifiable when binary covariates are subject to misclassification. To overcome model nonidentifiability, we consider scenarios where the misclassification probabilities are known, or the main/validation study design is available, and consequently, we propose estimation procedures for Markov models with binary covariates subject to misclassification. Simulation studies are conducted to evaluate the performance of the proposed methods and the consequence of the naive analysis which ignores the misclassification. Our proposed methods are illustrated by the application to the data arising from a psoriatic arthritic study.

Keywords and phrases: Identifiability, Main/validation study design, Markov models, Misclassification, Panel data.

1. INTRODUCTION

Misclassification of the variable arises commonly from many applications. In medical studies, for instance, diagnostic tests are the basic tools to measure the conditions for patients. However, they inevitably involve diagnostic errors and cannot perfectly reflect the accurate conditions of every subject. Misclassification may come from reading error induced from inaccurate machines or inexperience of readers. Sometimes, the accurate measurement of a variable is too costly or time-consuming to collect, and we have to take a surrogate measurement which is quick and cheap to obtain (e.g., Carroll, Gail and Lubin 1993).

Investigation of the impact of misclassification on analysis dates back to Bross (1954). An overview of the development was given by Kuha, Skinner and Palmgren (2005) who described the effects of misclassification and summarized the methods for adjusting misclassification effects. Given that misclassification parameters are estimated from validation studies or repeated measurements, consistent estimates of the relative risk and related parameters can be obtained from the matrix method (e.g., Bross 1954; Marshall 1990; Morrissey and Spiegelman 1999) or the maximum likelihood method (e.g., Espeland and Hui 1987). Küchenhoff, Mwalili and Lesaffre (2006) developed a simulation based method for parameter estimation in the presence of misclassification in discrete covariates, and Küchenhoff, Lederer and Lesaffre (2007) derived the asymptotic variance estimation for this approach. Yi et al. (2015) developed inference methods to address misclassified discrete covariates together with the presence of measurement error in continuous covariates. Various methods of handling data with misclassification were documented in the monographs including Gustafson (2004), Carroll et al. (2006), Buonaccorsi (2010), and Yi (2017).

Although there has been research on dealing with misclassified covariates, little attention has been paid to the covariate misclassification for the analysis of panel data under multi-state models. Multi-state models are useful tools for delineating the dynamic changes among different states of the response variable. When subjects are assessed periodically over a time period, exact transition times among the states are usually not observed, and only the state occupied at each assessment, together with the measurements of risk factors, is available. Such data are often called panel data (e.g., Kalbfleisch and Lawless 1985; Cook, Kalbfleisch and Yi 2002).

In analyzing panel data, Markov models are perhaps the most frequently used multi-state models due to their simplicity and interpretability. The Markov process is memoryless in that only the currently occupied state is relevant in specifying the transition intensities. Markov models have been studied by many authors to handle panel data under different settings. To name a few, see Kalbfleisch and Lawless (1985), Lindsey and Ryan (1993), Gentleman et al. (1994), Chen and Sen (1999), Hsieh, Chen and Chang (2002), Saint-Pierre et al. (2003), Cook, Zeng and Lee (2008), Hubbard, Inoue and Fann (2008), van den Hout and Matthews (2009), Chen, Yi and Cook (2010), and Tom and Farewell (2011), among many others.

In the presence of misclassified covariates, however, these methods break down. It is unclear what the impact of misclassification is on usual inferential procedures for analysis of panel data under Markov models. In this article, we investigate this important problem and consider Markov models with misclassified covariates. We examine a fundamental issue, model identifiability, which ubiquitously occurs in the

[*] Corresponding author.

presence of misclassified variables. To highlight the idea, the discussion is directed to binary covariates. We analytically show that misclassification in binary covariates breaks down the model identifiability which is well established for Markov models in the error-free context. To develop valid inference methods to account for misclassification effects, we consider three practical scenarios with: (1) known misclassification probabilities, (2) the main study/ internal validation design, and (3) the main study/external validation design. In the literature of measurement error models, having a validation subsample is a common requirement for characterizing the measurement error process. Main study/validation designs frequently appear in clinical trials and epidemiological studies; discussion and examples may be found in, for example, Greenland (1988), Willett (1998), Spiegelman, Rosner and Logan (2000), Spiegelman, Carroll and Kipnis (2001), Carroll et al. (2006), among many others.

This research is partially motivated by the data arising from a psoriatic arthritic study (Jackson 2011). Psoriatic arthritis (PsA) is a progressive disease for which the progression can be related to many factors. One interesting question is to understand how a binary risk factor concerning the information of effusions is associated with the disease progression. To answer this question, it is convenient to employ a progressive multi-state model to describe the transition intensities from state to state. To account for possible misclassification effects of the covariate, it is necessary to develop a valid estimation procedure for data analysis.

Although this work is motivated by the PsA data, the developed methods can be applied to general settings with Markov models in the presence of a misclassified covariate. This work complements the available research on Markov models with misclassified states (e.g., Rosychuk and Thompson 2003, 2004; Yi, He and He 2017).

The remainder is organized as follows. In Section 2, Markov models are described for the error-free context, and in Section 3, we show that the Markov models with misclassified binary covariates are not identifiable. The maximum likelihood estimation procedures are developed in Section 4, where either known misclassification probabilities or main study/validation study designs are considered. Simulation studies are conducted in Section 5 to demonstrate the performance of the proposed methods. Data arising from a psoriatic arthritic (PsA) study are analyzed using the proposed methods in Section 6. A general discussion is given in Section 7, and technical notes are presented in the appendix.

2. MODEL FORMULATION

2.1 Time-homogeneous Markov models

Suppose an individual moves among K states which are indexed by integers $1, \ldots, K$. Let $S(t)$ denote the state that is occupied by an individual at time t. Assume that $\{S(t) : t \geq 0\}$ follows a continuous-time Markov process. Let $P(s, s+t)$ be the $K \times K$ transition probability matrix from time s to time $s+t$ with entry (j, k) given by

$$p_{jk}(s, s+t) = P\{S(s+t) = k \mid S(s) = j\}$$

for $s \geq 0$, $t > 0$ and $j, k = 1, \ldots, K$. The transition intensity from state j to state k at time t is defined as

$$q_{jk}(t) = \lim_{\Delta t \downarrow 0} \frac{p_{jk}(t, t+\Delta t)}{\Delta t} \quad \text{for } j \neq k,$$

and as a convention, q_{jj} is defined as

$$q_{jj}(t) = -\sum_{k \neq j} q_{jk}(t).$$

Let $Q(t)$ be the $K \times K$ transition intensity matrix with entry (j, k) given by $q_{jk}(t)$, where $j, k = 1, \ldots, K$.

This article is primarily concerned with time-homogeneous Markov models which are often used to analyze panel data due to their simplicity (e.g., Kalbfleisch and Lawless 1985), where $q_{jk}(t)$ is assumed to be a constant for any t. We therefore write $q_{jk}(t) = q_{jk}$ for $j, k = 1, \ldots, K$ and $Q(t) = Q$, where q_{jk} is a nonnegative constant and Q is a $K \times K$ matrix with nonnegative elements. It follows that $P(s, s+t) = P(0, t)$, which is then written as $P(t)$ for ease of exposition.

For the time-homogeneous Markov model, transition probabilities are expressed in terms of transition intensities (Cox and Miller 1965, Chapter 4),

$$P(t) = \exp(Qt) = \sum_{l=0}^{\infty} Q^l \frac{t^l}{l!}, \tag{1}$$

where the matrix exponential is defined as the power series of the matrix product; and $Q^0 = I_K$, a $K \times K$ identity matrix.

To compute $P(t)$ using (1), one may use the matrix decomposition to work out a convenient algorithm. If Q has distinct eigenvalues, say, $d_1, \ldots, d_K$, then we write

$$Q = JDJ^{-1},$$

where $D = \operatorname{diag}(d_1, \ldots, d_K)$, and J is the $K \times K$ matrix whose jth column is the eigenvector associated with d_j. Then $P(t)$ is calculated as (Kalbfleisch and Lawless 1985):

$$P(t) = J\operatorname{diag}\{\exp(d_1 t), \ldots, \exp(d_K t)\} J^{-1}.$$

If Q has repeated eigenvalues, Kalbfleisch and Lawless (1985) suggested an analogous decomposition of Q to the Jordan canonical form (Cox and Miller 1965, Chapter 3). For practically useful models, see Chiang (1980) for detailed expressions.

2.2 Regression model with covariates incorporated

2.2.1 Transition intensity model

Let X be a binary covariate subject to misclassification and let x_1 and x_2 denote the two possible values that X assumes. Let Z stand for the vector of perfectly measured time-independent covariates. Given $\{X, Z\}$, $\{S(t) : t \geq 0\}$ is assumed to follow a time-homogeneous Markov model with the conditional transition intensity $q_{jk}(X, Z)$ for $j \neq k$ and $j, k = 1, \ldots, K$.

To model the effects of covariates on transitions, we consider the log-linear model for the intensity

$$q_{jk}(X, Z) = \exp(\beta_{jk0} + X\beta_{jkx} + Z^{\mathrm{T}}\beta_{jkz}) \tag{2}$$

for $j \neq k$ and $j, k = 1, ..., K$, where β_{jk0}, β_{jkx}, and β_{jkz} are the regression coefficients.

Let $\beta = (\beta_{jk0}, \beta_{jkx}, \beta_{jkz}^{\mathrm{T}} : j \neq k; j, k = 1, \ldots, K)^{\mathrm{T}}$, which is of our primary interest to estimate. The model (2) has been commonly used in applications; see, for example, Kalbfleisch and Lawless (1985) and Cook, Kalbfleisch and Yi (2002).

2.2.2 Misclassification model

Let X^* be a surrogate measurement of X. Often the non-differential misclassification mechanism is assumed with

$$P(S(t) = k | X^*, X, Z) = P(S(t) = k | X, Z)$$
$$\text{for any } t > 0 \text{ and } k = 1, \ldots, K.$$

This assumption says that the observed measurement X^* does not carry information about the outcome if the true covariates $\{X, Z\}$ are controlled (Carroll et al. 2006; Yi 2017).

For $l \neq r$ and $l, r = 1, 2$, let

$$\lambda_{lr}(Z) = P(X = x_r | X^* = x_l, Z)$$

be the misclassification probability of the true covariate X given the observed surrogate measurement X^* and precesely measured covariates Z (Yi 2017, Chapter 6). The logistic models are used to facilitate the effects of covariates on misclassification probabilities,

$$\begin{aligned} \log\left\{\frac{\lambda_{12}(Z)}{1 - \lambda_{12}(Z)}\right\} &= \alpha_{10} + Z^{\mathrm{T}}\alpha_{1z}; \\ \log\left\{\frac{\lambda_{21}(Z)}{1 - \lambda_{21}(Z)}\right\} &= \alpha_{20} + Z^{\mathrm{T}}\alpha_{2z} \end{aligned} \tag{3}$$

where $\alpha_{10}, \alpha_{20}, \alpha_{1z}$ and α_{2z} are regression coefficients. Write $\alpha = (\alpha_{l0}, \alpha_{lz}^{\mathrm{T}} : l = 1, 2)^{\mathrm{T}}$.

3. MODEL IDENTIFIABILITY

In this section, we show that given the model setup in Section 2.2, the joint model for the state process and the misclassification process is not identifiable.

Theorem:

Suppose that conditional on $\{X, Z\}$, $\{S(t) : t \geq 0\}$ is a Markov process which has the transition intensity matrix Q modeled by (2). Let $\mathbb{S} = (S_0, \ldots, S_m)^{\mathrm{T}}$ denote the states of the process $\{S(t) : t \geq 0\}$ observed at time points $0 = t_0 < t_1 < \ldots < t_m$ where $S_j = S(t_j)$ for $j = 0, \ldots, m$. Assume that the misclassification models are given by (3) and that the distribution of the initial state $Pr(S_1|X, Z)$ is free of the parameters of models (2) and (3).

For any given parameters α and β, we consider another set of parameters α^* and β^*, defined as

$$\begin{aligned} &\alpha^* = -\alpha \textit{ and} \\ &\beta^* = (\beta_{jk0}^*, \beta_{jkx}^*, \beta_{jkz}^{*\mathrm{T}} : j \neq k; j, k = 1, \ldots, K)^{\mathrm{T}}, \end{aligned}$$

where $\beta_{jk0}^* = \beta_{jk0} + \beta_{jkx}(x_1 + x_2), \beta_{jkx}^* = -\beta_{jkx}$, and $\beta_{jkz}^* = \beta_{jkz}$. Then the conditional probability of $\mathbb{S}$, given the observed covariates $\{X^*, Z\}$, cannot be differentiated at the two sets of parameter values $\{\alpha, \beta\}$ and $\{\alpha^*, \beta^*\}$, i.e.,

$$P(\mathbb{S}|X^*, Z; \alpha^*, \beta^*) = P(\mathbb{S}|X^*, Z; \alpha, \beta).$$

The proof of the theorem is included in . This theorem says that two distinct sets of parameters can lead to the same probability mass function of $\mathbb{S}$, thus, suggesting that the model is non-identifiable in the presence of misclassified binary covariates. Consequently, in developing valid inference methods to account for covariate misclassification effects, one needs to carefully address the issue of non-identifiability.

In the misclassification-free context, inference about the parameter β can be based on model (2) by using the likelihood method, and in this instance non-identifiability is not a concern. However, in the presence of variable misclassification, inferences often require additional modeling for the misclassification process, besides routine modeling the response process. This additional modeling of the nuisance process enlarges the initial parameter space, say Θ_β, for the response model (2) to a new parameter space, say $\Theta = \Theta_\beta \times \Theta_\alpha$, where Θ_α represents the space of the parameters induced from the additional modeling of the misclassification process. The new parameter space Θ is larger than the initial parameter space Θ_β in that additional dimensions $\dim(\Theta_\alpha)$ are resulted in, which creates possible model non-identifiability.

To overcome non-identifiability issues, we often impose certain constraints to the parameter space Θ to make it smaller, or equivalently, to make some parameter values inadmissible. A principle of imposing suitable constraints on the parameter space Θ is to preserve the initial model structures for the response process with the parameter set Θ_β unchanged but place constraints on the nuisance parameter set Θ_α.

A simple strategy is to assume that the values of the nuisance parameters (i.e., α) are known, say α_0, then the parameter space Θ becomes $\Theta_\beta \times \{\alpha_0\}$, which is essentially equivalent to the initial parameter space Θ_β for the response model (2). Assuming the nuisance parameters α to be known is a routine for conducting sensitivity analyses, where one typically specifies nuisance parameters to be representative values and then carries out inference about β to uncover how sensitive the results are to different magnitudes of nuisance parameters. This strategy is usually employed when the available data include only the surrogate measurements of the covariates together with the response measurements (e.g., Gustafson 2004; Carroll et al. 2006; Yi 2017).

In some applications, a validation subsample which contains the measurements for both the true covariaes and their surrogate measurements, together with the response measurements, is available (e.g., Spiegelman, Rosner and Logan 2000; Yi et al. 2019). In this case, we are able to estimate the nuisance parameter α using the validation data, and thus the parameter space Θ becomes $\Theta_\beta \times \{\hat{\alpha}\}$, where $\hat{\alpha}$ is the estimate of the parameter α. This virtually reduces to the preceding instance where model non-identifiability is not a problem.

In the next section we explore procedures for the estimation of the model parameter β for these scenarios.

4. MAXIMUM LIKELIHOOD METHODS

Suppose that there is a sample consisting of independent measurements of n individuals. We add subscript i to the symbols defined in the previous sections. For $i = 1, \ldots, n$, let $S_i(t)$ denote the state for individual i at time $t \geq 0$. Let X_i represent the true covariate, X_i^* denote the surrogate measure of X_i, and Z_i be the vector of precisely measured time-independent covariates for individual i. We assume that conditional on $\{X_i, Z_i\}$, $\{S_i(t) : t \geq 0\}$ follows a continuous time-homogeneous Markov process. Let $t_{i0} < t_{i1} < \ldots < t_{im_i}$ denote the $(m_i + 1)$ times at which individual i is observed. For simplicity, let S_{ij} denote $S_i(t_j)$ and $\mathbb{S}_i = (S_{i0}, \ldots, S_{im_i})^{\mathrm{T}}$.

Models described in the previous sections are employed to feature the transition process as well as the misclassification process. In order to estimate the parameters associated with the transition intensity model (2), we propose the likelihood inference methods for the situations discussed in Section 3: one is that the parameters in misclassification probabilities are known from empirical studies, and the other is that a validation sample is available together with the main study data.

4.1 Misclassification probabilities are known

First, we consider estimation procedures of β for the case where the parameter α in the misclassification models (3) is known as α_0, say. The likelihood function contributed from individual i is

$$
\begin{aligned}
(4) \qquad L_i(\beta) &= P(\mathbb{S}_i | X_i^*, Z_i; \alpha_0, \beta) \\
&\propto P(X_i = x_1 | X_i^*, Z_i; \alpha_0) \\
&\quad \cdot \prod_{j=1}^{m_i} P(S_{ij} | S_{i,j-1}, X_i = x_1, Z_i; \beta) \\
&\quad + P(X_i = x_2 | X_i^*, Z_i; \alpha_0) \\
&\quad \cdot \prod_{j=1}^{m_i} P(S_{ij} | S_{i,j-1}, X_i = x_2, Z_i; \beta),
\end{aligned}
$$

where for $r = 1, 2$, $P(X_i = x_r | X_i^*, Z_i; \alpha_0)$ is determined by (3), $P(S_{ij} | S_{i,j-1}, X_i = x_r, Z_i; \beta)$ is the transition probability defined in Section 2.2.1, and $P(S_{i0} | X_i, Z_i)$ is the initial state occupation probability which is assumed to be free of parameter β.

Thus, the likelihood function of β is

$$L(\beta) = \prod_{i=1}^{n} L_i(\beta).$$

The maximum likelihood estimator of β, denoted by $\hat{\beta}$, can be obtained by maximizing the log-likelihood $\log L(\beta)$ with respect to β. To implement this, one may employ the Newton-Raphson algorithm described by Kosorok and Chao (1996) or the quasi-Newton algorithm proposed by Kalbfleisch and Lawless (1985).

From standard likelihood theory, under regularity conditions (e.g., Andersen et al. 1993, Section 8.3; Lehmann 1999, Section 7.3), the maximum likelihood estimator $\hat{\beta}$ is a consistent estimator of β and has an asymptotic normal distribution given by

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Sigma^{-1}(\beta)) \qquad \text{as } n \to \infty,$$

where $\Sigma(\beta) = E\left[\{\partial \log L_i(\beta)/\partial\beta\}^{\bigotimes 2}\right]$, and $a^{\bigotimes 2} = aa^{\mathrm{T}}$ for a column vector a.

When using this asymptotic distribution to conduct inference about β, $\Sigma(\beta)$ is replaced by its consistent estimate $\hat{\Sigma}(\beta) = n^{-1} \sum_{i=1}^{n} \{\partial \log L_i(\beta)/\partial\beta\}^{\bigotimes 2}|_{\beta=\hat{\beta}}$, where

$$
\begin{aligned}
\frac{\partial \log L_i(\beta)}{\partial \beta} &= \frac{1}{L_i(\beta)} \Bigg(\sum_{r=1,2} \Bigg[P(X_i = x_r | X_i^*, Z_i; \alpha_0) \\
&\quad \cdot \sum_{j=1}^{m_i} \Bigg\{ \frac{\partial P(S_{ij} | S_{i,j-1}, X_i = x_r, Z_i; \beta)}{\partial \beta} \\
&\quad \cdot \prod_{1 \leq l \neq j \leq m_i} P(S_{il} | S_{i,l-1}, X_i = x_r, Z_i; \beta) \Bigg\} \Bigg] \Bigg).
\end{aligned}
$$

4.2 Main study/validation study

In applications, parameter α in the misclassification models may be unknown and must be estimated from additional data sources. Here we describe an estimation method when a validation sample is available in addition to the main study data. Our development covers two types of validation studies: *internal* or *external*, which are used in estimation of parameter $\theta = (\beta^{\mathrm{T}}, \alpha^{\mathrm{T}})^{\mathrm{T}}$, together with the main study.

Main Study/Internal Validation Design

In the main study/internal validation design, in addition to the main study data $\{(X_i^*, Z_i, \mathbb{S}_i) : i \in \mathcal{M}\}$, an internal validation study with data $\{(X_i, X_i^*, Z_i, \mathbb{S}_i) : i \in \mathcal{V}\}$ is available, where $\mathcal{M}$ and $\mathcal{V}$ are the index sets for the main study and validation study, respectively, and $\mathcal{V}$ is a subset of $\mathcal{M}$. We assume that the validation sample $\mathcal{V}$ is a random subsample of the main study sample. Let δ_i denote the selection indicator for individual i, where $\delta_i = 1$ if individual i is selected to be included in the validation sample and $\delta_i = 0$ otherwise. We let n_{V} represent the size of $\mathcal{V}$ in contrast to the size n of $\mathcal{M}$. In circumstances where measuring the true value of X_i is expensive or time-consuming, n_{V} is typically much smaller than n.

For $i \in \mathcal{V}$, the likelihood contributed from individual i is given by

$$\begin{aligned} L_{\mathrm{V}i} &= P(\delta_i = 1, X_i, \mathbb{S}_i | X_i^*, Z_i; \alpha, \beta) \\ &\propto P(\mathbb{S}_i | X_i, Z_i; \beta) P(X_i | X_i^*, Z_i; \alpha), \end{aligned}$$

and for $i \in \mathcal{M} \backslash \mathcal{V}$, the likelihood contributed from individual i is given by

$$\begin{aligned} L_{\mathrm{M}i} &= P(\delta_i = 0, \mathbb{S}_i | X_i^*, Z_i;\ \alpha, \beta) \\ &\propto P(\mathbb{S}_i | X_i = x_1, Z_i; \beta) P(X_i = x_1 | X_i^*, Z_i; \alpha) \\ &\quad + P(\mathbb{S}_i | X_i = x_2, Z_i; \beta) P(X_i = x_2 | X_i^*, Z_i; \alpha), \end{aligned} \tag{5}$$

where $P(\delta_i = r | X_i, X_i^*, Z_i)$ is the probability for including or not including subject i in the validation study which is assumed to satisfy

$$P(\delta_i = r | X_i, X_i^*, Z_i) = P(\delta_i = r | Z_i)$$

for $r = 0, 1$, i.e., the selection of individual i into the validation study does not depend on either the true X_i or observed X_i^*, given perfectly measured covariates Z_i.

Consequently, the likelihood for a main study/internal validation study design is

$$L_{int}(\alpha, \beta) = \prod_{i \in \mathcal{V}} L_{\mathrm{V}i} \cdot \prod_{i \in \mathcal{M} \backslash \mathcal{V}} L_{\mathrm{M}i}.$$

Therefore, the log-likelihood for the main study/internal validation study design takes the form

$$\begin{aligned} \log L_{int}(\alpha, \beta) &= \sum_{i \in \mathcal{M} \backslash \mathcal{V}} \log \{P(\mathbb{S}_i | X_i^*, Z_i; \alpha, \beta)\} \\ &\quad + \sum_{i \in \mathcal{V}} \log \{P(X_i | X_i^*, Z_i; \alpha)\} \\ &\quad + \sum_{i \in \mathcal{V}} \sum_{j=1}^{m_i} \log \{P(S_{ij} | S_{i,j-1}, X_i, Z_i; \beta)\}, \end{aligned}$$

where $P(\mathbb{S}_i | X_i^*, Z_i; \alpha, \beta)$ is given by (5) with α_0 replaced by α, $P(X_i | X_i^*, Z_i; \alpha)$ is the misclassification probability determined by (3), and $P(S_{ij} | S_{i,j-1}, X_i, Z_i; \beta)$ is the transition probability defined in Section 2.2.1. The maximum likelihood estimates of α and β can be obtained by maximizing the log-likelihood $\log L_{int}(\alpha, \beta)$ with respect to α and β. Let $\hat{\theta}_{int}$ denote the resultant estimator of θ.

Let $S_{\mathrm{V}i\alpha} = \partial \log L_{\mathrm{V}i} / \partial \alpha$, $S_{\mathrm{V}i\beta} = \partial \log L_{\mathrm{V}i} / \partial \beta$, $S_{\mathrm{M}i\alpha} = \partial \log L_{\mathrm{M}i} / \partial \alpha$, $S_{\mathrm{M}i\alpha} = \partial \log L_{\mathrm{M}i} / \partial \beta$, and $S_{\mathrm{M}i\theta} = (S_{\mathrm{M}i\beta}^{\mathrm{T}}, S_{\mathrm{M}i\alpha}^{\mathrm{T}})^{\mathrm{T}}$. Under regularity conditions and when the ratio n_{V}/n approaches a positive constant ρ as $n \to \infty$, $\sqrt{n}(\hat{\theta}_{int} - \theta)$ has an asymptotic normal distribution with mean zero and covariance matrix A_{int}^{-1}, where

$$\begin{aligned} A_{int} &= -(1-\rho) E\left(\frac{\partial S_{\mathrm{M}i\theta}}{\partial \theta^{\mathrm{T}}}\right) \\ &\quad - \rho \mathrm{diag}\left\{E\left(\frac{\partial S_{\mathrm{V}i\beta}}{\partial \beta^{\mathrm{T}}}\right), E\left(\frac{\partial S_{\mathrm{V}i\alpha}}{\partial \alpha^{\mathrm{T}}}\right)\right\}. \end{aligned}$$

This result can be proved by modifying standard likelihood theory; a sketch is given in Appendix B.

Main Study/External Validation Design

In the main study/external validation design, the available data are $\{(X_i^*, Z_i, \mathbb{S}_i) : i \in \mathcal{M}\}$ and $\{(X_i^*, X_i, Z_i) : i \in \mathcal{V}\}$, respectively, where $\mathcal{V}$ and $\mathcal{M}$ do not overlap, and there are no response measurements for subjects in $\mathcal{V}$. We still use n and n_{V} to denote the size of $\mathcal{M}$ and $\mathcal{V}$, respectively.

With the main study/external validation design, we assume that given Z_i, the conditional distribution of (X_i, X_i^*) for $i \in \mathcal{V}$ is the same as that of (X_i, X_i^*) for $i \in \mathcal{M}$ so that the information carried by the study $\mathcal{V}$ can be transported to the main study $\mathcal{M}$ when carrying out inferences. The feasibility of this assumption is justified by subject matter considerations. This assumption is typically reasonable for scenarios where both main and external validation studies are carried out to the same population using the same data collection procedures (e.g., Yi et al. 2015).

For $i \in \mathcal{V}$, the likelihood contributed from individual i is given by $P(\delta_i = 1, X_i, | X_i^*, Z_i; \alpha)$, and for $i \in \mathcal{M}$, the likelihood contributed from individual i is given by (5). Then the log-likelihood for the main study/external validation study design is given by

$$\begin{aligned} \log L_{ext}(\alpha, \beta) &= \sum_{i \in \mathcal{M}} \log [P(\mathbb{S}_i | X_i^*, Z_i; \alpha, \beta)] \\ &\quad + \sum_{i \in \mathcal{V}} \log [P(X_i | X_i^*, Z_i; \alpha)]. \end{aligned}$$

Maximizing $\log L_{ext}(\alpha, \beta)$ with respect to α and β yields estimator of α and β. Let $\hat{\theta}_{ext}$ be the resulting estimator of θ. Under regularity conditions and when the ratio n_{V}/n approaches a positive constant ρ as $n \to \infty$,

$$\sqrt{n}(\hat{\theta}_{ext} - \theta) \xrightarrow{d} N\left(0, \frac{1}{1+\rho} A_{ext}^{-1}\right) \quad \text{as} \quad n \to \infty,$$

Table 1. Simulation results for three-state progressive models with a misclassified binary covariate: misclassification probabilities are given

	TRUE				NAIVE							
	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%				
β_{10}	.005	.047	.047	94.6	-.045	.050	.050	85.4				
β_{1x}	-.001	.047	.045	95.4	.085	.050	.051	59.7				
β_{1z}	.004	.049	.048	95.3	.000	.050	.051	94.5				
β_{20}	.003	.051	.053	93.2	-.071	.054	.055	73.6				
β_{2x}	-.002	.051	.050	95.2	.130	.054	.054	33.6				
β_{2z}	.003	.054	.057	93.6	-.014	.056	.056	93.8				
	Case 1: $(\alpha_1, \alpha_2) = (0.3, 0.1)$				Case 2: $(\alpha_1, \alpha_2) = (0.2, 0.05)$				Case 3: $(\alpha_1, \alpha_2) = (0.4, 0.15)$			
	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%
β_{10}	.006	.049	.050	93.7	-.015	.048	.048	92.7	.032	.052	.054	90.5
β_{1x}	.004	.079	.080	92.9	.038	.069	.070	90.0	-.027	.089	.089	91.7
β_{1z}	.004	.050	.051	94.8	.002	.050	.050	95.0	.004	.051	.051	94.9
β_{20}	.005	.054	.053	94.9	-.027	.052	.052	91.5	.046	.058	.058	89.7
β_{2x}	.003	.084	.082	95.0	.052	.077	.073	89.7	-.037	.089	.087	92.7
β_{2z}	.002	.057	.059	93.5	-.005	.057	.059	93.0	.006	.058	.060	93.7

"Bias" represents the difference between the parameter value and its estimate, "ASE" refers to the model-based asymptotic standard error, "ESE" stands for the empirical standard deviation, and "CR%" displays the coverage rate (in percentage) for 95% confidence intervals.

where

$$A_{ext} = -\frac{1}{(1+\rho)} E\left(\frac{\partial S_{\mathrm{M}i\theta}}{\partial \theta^{\mathrm{T}}}\right) - \frac{\rho}{1+\rho} \mathrm{diag}\left\{0, E\left(\frac{\partial S^*_{\mathrm{V}i\alpha}}{\partial \alpha^{\mathrm{T}}}\right)\right\}$$

with $S^*_{\mathrm{V}i\alpha} = \partial \log P(\delta_i = 1, X_i | X_i^*, Z_i; \alpha)/\partial\alpha$. Here and elsewhere, the symbol 0 may represent the number zero as well as a zero vector or a zero matrix whose dimension is clear from the context. A derivation of this result is outlined in Appendix C.

5. SIMULATION STUDIES

In this section we carry out simulation studies to evaluate the performance of the proposed methods and demonstrate the consequence of the naive method which ignores the covariate misclassification. The sample size is $n = 500$ and 1,000 simulations are run for each parameter configuration.

We consider a three-state progressive time-homogenous Markov model where the transition intensity is given by

$$q_{j,j+1} = \exp(\beta_{j0} + X\beta_{jx} + Z\beta_{jz}) \tag{6}$$

for $j = 1, 2$, where we set $\beta_{10} = -1.0, \beta_{1x} = -0.2, \beta_{1z} = 0.6, \beta_{20} = -0.7, \beta_{2x} = -0.3$, and $\beta_{2z} = 0.5$.

Each individual is assumed to start from state 1 at the initial time $t_0 = 0$ and is observed at the examination times, $t_1 < \ldots < t_m$, where m is taken as 11. The gap between two adjacent examination times, $t_{j+1} - t_j$, is uniformly distributed on the interval [0.5; 1.0], where $j = 0, \ldots, m-1$. A continuous covariate Z is generated from the standard normal distribution. The observed binary covariate X^* independently takes values -1 and 1 with the respective probabilities 2/3 and 1/3. Conditional on X^*, the true binary covariate X is independent of Z and is generated based on the misclassification probabilities

$$\begin{aligned}\alpha_1 &= P(X = 1 | X^* = -1) = 0.3,\\ \alpha_2 &= P(X = -1 | X^* = 1) = 0.1.\end{aligned}$$

We analyze the simulated data using the methods developed in Section 4. First, we assume that the misclassification probabilities are known and we particularly consider the following three scenarios:

Case 1: The misclassification probabilities are specified as the values used for data generation, i.e., $(\alpha_1, \alpha_2) = (0.3, 0.1)$;

Case 2: The misclassification probabilities are specified to be smaller than the true values and we take $(\alpha_1, \alpha_2) = (0.2, 0.05)$;

Case 3: The misclassification probabilities are specified to be larger than the true values and we take $(\alpha_1, \alpha_2) = (0.4, 0.15)$.

Case 1 is a reflection of the true scenario of the misclassification process, and Cases 2 and 3 facilitate circumstances where the misclassification probabilities are misspecified. For comparison purposes, we also run two analyses. In the first analysis (called "TRUE"), we pretend the true covariate measurements X_i are available and use them to fit the model (6); and in the second analysis (called "NAIVE"), we perform the naive analysis by fitting model (6) with X_i directly replaced by the observed measurements X_i^*.

The analysis results are reported in Table 1 where "Bias" represents the difference between the parameter value and its estimate, "ASE" refers to the model-based asymptotic standard error, "ESE" stands for the empirical standard deviation, and "CR%" displays the coverage rate (in percentage) for 95% confidence intervals.

Table 2. Simulation results for three-state progressive models with a misclassified binary covariate: the main/validation study is considered

	Internal validation study							
	$n_2 = 50$				$n_2 = 100$			
	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%
β_{10}	.006	.057	.052	96.5	.006	.049	.049	95.4
β_{1x}	.003	.076	.073	94.3	-.000	.067	.066	95.1
β_{1z}	.003	.048	.050	94.8	.001	.046	.046	94.2
β_{20}	.007	.071	.062	97.2	.004	.058	.052	96.2
β_{2x}	-.003	.083	.079	94.9	-.001	.071	.068	95.3
β_{2z}	.000	.055	.055	94.9	.002	.052	.052	94.9
α_{10}	-.034	.539	.407	98.7	-.016	.395	.276	99.2
α_{20}	-.098	1.018	.574	100.0	-.062	.617	.439	100.0

	External validation study							
	$n_2 = 50$				$n_2 = 100$			
	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%
β_{10}	.004	.064	.057	95.8	.009	.055	.053	95.8
β_{1x}	.003	.088	.083	93.8	.002	.082	.084	93.3
β_{1z}	.004	.051	.051	94.8	.004	.050	.051	94.1
β_{20}	.004	.080	.065	97.4	.006	.065	.064	95.4
β_{2x}	.000	.096	.087	94.2	.003	.089	.086	94.9
β_{2z}	.000	.058	.060	93.3	.001	.058	.061	94.0
α_{10}	-.086	.502	.425	98.0	-.010	.330	.295	97.4
α_{20}	-.091	1.174	.604	100.0	-.109	.633	.457	99.8

"Bias" represents the difference between the parameter value and its estimate, "ASE" refers to the model-based asymptotic standard error, "ESE" stands for the empirical standard deviation, and "CR%" displays the coverage rate (in percentage) for 95% confidence intervals.

As expected, the analysis with the true measurements of X_i used gives the best estimation results among all the analysis methods. Using these results as a reference point, we now examine the results obtained from other methods. The NAIVE method which disregards the feature of misclassification produces noticeably biased results; the finite sample biases are large and the coverage rates of 95% confidence intervals deviate from the nominal level for the intercepts and the X_i covariate effects. On the other hand, the proposed method under Case 1 significantly outperforms the NAIVE method. It yields results that are quite close to those produced by the TRUE method. Finite sample biases are reasonably small, the model-based variance estimates agree well with the empirical variances, and the coverage rates for 95% confidence intervals are close to the nominal level. Unsurprisingly, when misclassification probabilities are misspecified, the performance of the proposed method (i.e., under Cases 2 and 3) would deteriorate and biased results may incur. Interestingly, in the cases we consider here, the proposed method still outperforms the NAIVE method even when misspecification of the misclassification probabilities is involved.

Next, we assess the performance of the proposed method when misclassification probabilities must be estimated from a validation sample. To form a validation sample, we randomly include n_{V} individuals such that half of them have $X_i = -1$ and half of them have $X_i = 1$, where n_{V} is 50 or 100. We consider the two cases where either an internal or an external validation sample is available in addition to the main study data, and apply the estimation methods described in Section 4.2 to estimate the parameter β.

The analysis results are recorded in Table 2 where the entries have the same meaning as for Table 1. Regarding estimation of the parameter β, the proposed method performs well for both circumstances with either an internal or an external sample. Finite sample biases are negligible, model-based asymptotic standard errors fairly agree with the empirical standard deviations, and the coverage rates for 95% confidence intervals are comparable to the nominal level. As expected, when the size in the validation sample increases, ASE and ESE of the estimators tend to decrease. Regarding estimation of the nuisance parameters α_{10} and α_{20}, we observe noticeable finite sample biases and the departure of the the coverage rates of the 95% confidence intervals from the nominal level. This is mainly owing to the small sizes of the validation samples. However, such unideal results do not seem to drastically affect consistent estimation of the parameter β which is of principal interest.

In summary, the simulation study demonstrates that the naive analysis with misclassification ignored yields biased estimation results. It is useful to adjust for misclassification effects in inferential procedures. The finite sample performance of the proposed methods is fairly satisfactory, which is indicated by the results we obtain here.

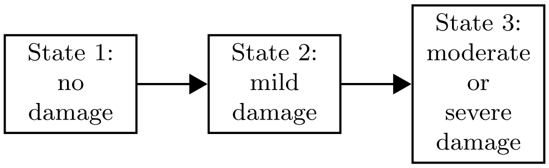

Figure 1. Three-state progressive model for the PsA study.

6. APPLICATION TO THE PSA DATA

To illustrate the proposed methods, we analyze the data arising from a psoriatic arthritic study (Jackson 2011). Psoriatic arthritis (PsA) is a progressive disease for which the progression is usually reflected in the accumulation and severity of damaged joints. Here we consider a three state progressive model, shown in Figure 1, to facilitate the progression of PsA: State 1 represents that a subject has no damaged joints, State 2 indicates a mild damage condition for which an individual has 1 to 4 damaged joints, and State 3 features a moderate or severe damage status of a subject who has 5 or more damaged joints. The data set, denoted by $\{(X_i, \mathbb{S}_i) : i = 1, \ldots, n\}$, contains $n = 305$ subjects with 806 observations which are obtained from the visits to a psoriatic arthritis clinic. The risk factor X_i is taken as the presence or absence of five or more effusions (coded as 'hieff': -1 for "no presence" and +1 for "presence"). This covariate is time-independent with 48 positive values and 257 negative values among all the subjects.

In the three-state progressive model, transition intensities $q_{j,j+1}$ are modeled by the log-linear model

$$\log\ q_{j,j+1} = \beta_{j0} + X_j \beta_{jx} \tag{7}$$

for $j = 1, 2$, where β_{j0} and β_{jx} are the regression coefficients to be estimated for $j = 1, 2$.

To see how misclassification may impact the analysis, we consider a scenario where the surrogate measurement, denoted by X_i^*, is available but X_i is not observed, and the surrogate measurement is related to the true covariate X_i in such a way that only one type of misclassification is present, i.e., $P(X_i = +1 | X_i^* = -1) = 0$ and $P(X_i = -1 | X_i^* = +1) > 0$. In particular, the surrogate measurement X_i^* is generated from the conditional probability mass function $P(X_i^* = +1 | X_i = +1) = P(X_i^* = -1 | X_i = -1) = 0.8$. If the generated value of X_i^* is -1, then we replace it with the value of X_i, i.e., set $X_i^* = X_i$.

To analyze the PsA data from different perspectives, we conduct the following four analyses.

Analysis 1:

The three-state progressive Markov model (7) is fitted to the PsA data $\{(X_i, \mathbb{S}_i) : i = 1, \ldots, n\}$.

Analysis 2:

We fit the three-state progressive Markov model (7) to the data $\{(X_i^*, \mathbb{S}_i) : i = 1, \ldots, n\}$, with X_i replaced by X_i^*. This is a naive method which ignores the misclassification.

Analysis 3:

We fit the three-state progressive Markov model (7) to the data $\{(X_i^*, \mathbb{S}_i) : i = 1, \ldots, n\}$ using the method described in Section 4.1, where the misclassification probability is reparameterized as

$$P(X_i = -1 | X_i^* = +1) = \frac{\exp(\alpha)}{1 + \exp(\alpha)}$$

with the parameter α taken as 0.5.

Analysis 4:

We fit the three-state progressive Markov model (7) to the data $\{(X_i^*, \mathbb{S}_i) : i = 1, \ldots, n\}$ using the method described in Section 4.2 for the main/internal validation data, where the main study contains the measurements $\{(X_i^*, \mathbb{S}_i) : i = 1, \ldots, n\}$ and the internal validation sample includes the measurements $\{(X_i, X_i^*, \mathbb{S}_i) : i = 1, \ldots, 30\}$ of 30 randomly selected subjects with a positive surrogate measurement X_i^*.

The analysis results are summarized in Table 3 where "EST" stands for the point estimate of a parameter, "ASE" refers to the model-based standard error of the associated estimator, and "*p*-value" records the p-value for the corresponding null hypothesis of no effect. Comparing the results obtained from the different analyses, we have the following findings.

Analysis 1 vs Analysis 2:

The point estimates and standard errors obtained from Analyses 1 and 2 are close for estimation of β_{10}, β_{20} and β_{2x}. The estimate of β_{1x} obtained from Analysis 2 is attenuated relative to that obtained from Analysis 1. The significant effect of the covariate X_i on the onset of PsA (States $1 \to 2$) is detected in Analysis 1 but not in Analysis 2, which reveals the consequence of ignoring the misclassification in Analysis 2.

Analysis 3 vs Analysis 4:

The point estimates in Analyses 3 and 4 agree well, and standard errors for the estimators of the parameters related to the disease progression (States $2 \to 3$) in both analyses are close. However, standard errors of $\hat{\beta}_{10}$ and $\hat{\beta}_{1x}$ in Analysis 3 are much larger than those obtained from Analysis 4 where $\hat{\beta}_{10}$ and $\hat{\beta}_{1x}$ represent the estimator of β_{10} and β_{1x}, respectively. The inflated standard error for $\hat{\beta}_{1x}$ may result in the failure of detecting the significant effect of the covariate hieff on the onset of PsA (States $1 \to 2$) in Analysis 3.

Analysis 4 vs Analysis 1:

The results obtained from the main study/internal validation study design (Analysis 4) agree fairly well with the results obtained using the true covariate (Analysis 1). Both methods successfully capture the significant effect of hieff on the onset of PsA (States $1 \to 2$) and give comparable estimates and p-values for all the parameters. Positive estimates of the covariate effect for hieff from the two analyses show that hieff has a positive effect on transition from one state to another; the presence of five or more effusions is likely to increase the transition rate. In addition, comparing the

Table 3. Analyses of PsA data under the three-state progressive model

			Analysis 1			Analysis 2		
	Covariate		EST	ASE	p-value	EST	ASE	p-value
Transition								
State 1 → 2	Intercept	β_{10}	-2.05	0.20	$< .001$	-2.14	0.21	$< .001$
	hieff	β_{1x}	0.42	0.20	.036	0.29	0.22	.169
State 2 → 3	Intercept	β_{20}	-1.71	0.16	$< .001$	-1.70	0.17	$< .001$
	hieff	β_{2x}	0.23	0.16	.135	0.25	0.17	.148

			Analysis 3			Analysis 4		
	Covariate		EST	ASE	p-value	EST	ASE	p-value
Transition								
State 1 → 2	Intercept	β_{10}	-1.87	0.72	0.010	1.93	0.27	$< .001$
	hieff	β_{1x}	0.68	0.84	.418	0.58	0.29	.045
State 2 → 3	Intercept	β_{20}	-1.62	0.23	$< .001$	-1.62	0.21	$< .001$
	hieff	β_{2x}	0.38	0.28	.177	0.37	0.23	.113

"EST" stands for the point estimate of a parameter, "ASE" refers to the model-based standard error of the associated estimator, and "p-value" records the p-value for the corresponding null hypothesis of no effect.

estimates of the covariate effect hieff for different transition intensities, we see that the presence of five or more effusions has a larger effect on increasing transition from states 1 to 2 than that from states 2 to 3.

In summary, taking the results of Analysis 1 as reference values, we see that Analysis 2 reveals the attenuated effects of the naive analysis which ignores the misclassification feature in the analysis. This also demonstrates the necessity of accounting for misclassification effects in the analysis of misclassification-prone data. Both Analysis 3 and Analysis 4 are developed to incorporate misclassification effects but they are applicable to different circumstances. In our current setting of surrogate data generation, it is expected that Analysis 4 produces the closest results to those of Analysis 1.

Finally, as noted by the Associate Editor, the data analysis we consider here solely serves for illustrative purposes. Because the initial dataset does not contain misclassification nor a validation sample, we artificially generate a subset of surrogate measurements X^* for X and then compare the performance of different analysis methods. Such studies simply demonstrate that *ignoring* or *incorporating* the feature of misclassification can yield different results, and when misclassification is involved, it is generally necessary to account for its effects on data analysis. Other than being illustrative, the analysis results here should not be over-interpreted for extracting new knowledge.

7. DISCUSSION

In this article, we investigate the problem of panel data with misclassified covariates. We study the impact of misclassification of a binary covariate on the model structure and analytically show the model non-identifiability. This result complements the discussion by Rosychuk and Thompson (2004) who explored the case with a binary outcome subject to misclassification. Our development in Section 3 capitalizes on the unique feature of the logistic regression for the binary variable, as shown by (9) in the appendix. The non-identifiability established in the theorem reflects the "symmetry" of the logistic model for a binary variable in the sense that the values of the variable are interchangeable if the model parameter takes values negative to each other. If the misclassification probabilities are not modeled by logistic regression models but are characterized by other regression forms, or if the misclassification-prone covariate X assumes more than two values, it is not obvious to prove or disprove non-identifiability in such settings. This is an interesting problem that warrants further explorations.

To address misclassification effects, in this paper we develop valid estimation procedures to analyze panel data with misclassified binary covariates. To overcome the non-identifiability problem, we propose the likelihood methods to make statistical inference and ensure the model identifiability in practical situations: one is based on the known misclassification probabilities, which is useful for conducting sensitivity analysis; the other one is developed based on the main study/validation study design. Simulation studies demonstrate satisfactory performance of our proposed methods.

Our methods are flexible for handling problems with unequally spaced assessment times, and moreover, the exact transition times are interval censored under the panel/intermittent observation scheme. We utilize continuous-time Markov models for analysis of panel data, and we are interested in understanding the influence of covariates on transitions among the states. Our methods extend the scope of existing approaches of dealing with panel data to accommodating a practical feature that covariates are mismeasured.

APPENDIX A: PROOF OF THE THEOREM

To show how different parameter values may affect modeling of the transition intensities $q_{jk}(X, Z)$ and the misclassification probabilities $\lambda_{lr}(Z)$, respectively, through model

(2) and model (3), rather than writing the right-hand-side of (2) and (3), we let $q_{jk}(X, Z; \beta)$ and $\lambda_{lr}(Z; \alpha)$ denote the the right-hand-side of (2) and (3), respectively, in the following derivations for ease of exposition.

First, we examine the misclassification model (3) under different parameter values α and α^*. Noting that $\alpha^* = -\alpha$ yields

$$\frac{\exp(\alpha_{l0} + Z^{\mathrm{T}}\alpha_{lz})}{1 + \exp(\alpha_{l0} + Z^{\mathrm{T}}\alpha_{lz})} = \frac{1}{1 + \exp(\alpha^*_{l0} + Z^{\mathrm{T}}\alpha^*_{lz})} \tag{8}$$

for any value of Z and $l = 1, 2$, we obtain that for $l, r = 1, 2$ with $l \neq r$, by (3) and (8),

$$\begin{aligned} \lambda_{lr}(Z; \alpha) &= 1 - \lambda_{lr}(Z; \alpha^*) \\ &= \lambda_{rr}(Z; \alpha^*). \end{aligned} \tag{9}$$

Thus we have

$$P(X = x_1 | X^* = x_2, Z; \alpha) = P(X = x_2 | X^* = x_2, Z; \alpha^*),$$

or more generally,

$$P(X = x_1 | X^*, Z; \alpha) = P(X = x_2 | X^*, Z; \alpha^*). \tag{10}$$

Next, we examine the response model (2) under different parameter values β and β^*. By the definition of β^*_{ij0}, β^*_{ijx} and β^*_{ijz}, we obtain that

$$\begin{aligned} \beta^*_{jk0} + \beta^*_{jkx}x_2 + Z^{\mathrm{T}}\beta^*_{jkz} &= \beta_{jk0} + \beta_{jkx}x_1 + Z^{\mathrm{T}}\beta_{jkz}; \\ \beta^*_{jk0} + \beta^*_{jkx}x_1 + Z^{\mathrm{T}}\beta^*_{jkz} &= \beta_{jk0} + \beta_{jkx}x_2 + Z^{\mathrm{T}}\beta_{jkz}; \end{aligned}$$

and hence, by (2),

$$\begin{aligned} & q_{jk}(x_1, Z; \beta) = q_{jk}(x_2, Z; \beta^*) \text{ and} \\ & q_{jk}(x_2, Z; \beta) = q_{jk}(x_1, Z; \beta^*). \end{aligned} \tag{11}$$

By (1), transition probabilities for the time-homogeneous Markov model are determined by transition intensities. Therefore, applying (11) and the assumption that $P(S_1 | X^*, Z)$ does not contain α or β, we obtain that

$$\begin{aligned} P(\mathbb{S} | X = x_1, Z; \beta) &= P(\mathbb{S} | X = x_2, Z; \beta^*); \\ P(\mathbb{S} | X = x_2, Z; \beta) &= P(\mathbb{S} | X = x_1, Z; \beta^*). \end{aligned} \tag{12}$$

As a result, we obtain

$$\begin{aligned} & P(\mathbb{S} | X^*, Z; \alpha, \beta) \\ = \; & P(\mathbb{S}, X = x_1 | X^*, Z; \alpha, \beta) \\ & + P(\mathbb{S}, X = x_2 | X^*, Z; \alpha, \beta) \\ = \; & P(\mathbb{S} | X = x_1, Z; \beta) P(X = x_1 | X^*, Z; \alpha) \\ & + P(\mathbb{S} | X = x_2, Z; \beta) P(X = x_2 | X^*, Z; \alpha) \\ = \; & P(\mathbb{S} | X = x_2, Z; \beta^*) P(X = x_2 | X^*, Z; \alpha^*) \\ & + P(\mathbb{S} | X = x_1, Z; \beta^*) P(X = x_1 | X^*, Z; \alpha^*) \\ = \; & P(\mathbb{S}, X = x_2 | X^*, Z; \alpha^*, \beta^*) \\ & + P(\mathbb{S}, X = x_1 | X^*, Z; \alpha^*, \beta^*) \\ = \; & P(\mathbb{S} | X^*, Z; \alpha^*, \beta^*), \end{aligned}$$

where the nondifferential misclassification mechanism is used in the second and fourth steps, and the third step is due to (10) and (12).

APPENDIX B

Under regularity conditions, maximizing $\log L_{int}(\alpha, \beta)$ is equivalent to solving

$$\sum_{i \in \mathcal{M}} \begin{pmatrix} (1 - \delta_i) S_{\mathrm{M}i\beta} + \delta_i S_{\mathrm{V}i\beta} \\ (1 - \delta_i) S_{\mathrm{M}i\alpha} + \delta_i S_{\mathrm{V}i\alpha} \end{pmatrix} = 0.$$

Let $H_{i\beta} = (1 - \delta_i) S_{\mathrm{M}i\beta} + \delta_i S_{\mathrm{V}i\beta}$, $H_{i\alpha} = (1 - \delta_i) S_{\mathrm{M}i\alpha} + \delta_i S_{\mathrm{V}i\alpha}$ and $H_{i\theta} = (H^{\mathrm{T}}_{i\beta}, H^{\mathrm{T}}_{i\alpha})^{\mathrm{T}}$. Then solving $\sum_{i \in \mathcal{M}} H_{i\theta} = 0$ yields the estimator $\hat{\theta}_{int}$. Applying the Taylor series expansion to $\sum_{i \in \mathcal{M}} H_{i\theta}(\hat{\theta}_{int}) = 0$ gives

$$\sum_{i \in \mathcal{M}} H_{i\theta} + \sum_{i \in \mathcal{M}} \frac{\partial H_{i\theta}}{\partial \theta^{\mathrm{T}}} (\hat{\theta}_{int} - \theta) + O_p(1) = 0,$$

leading to

$$\sqrt{n}(\hat{\theta}_{int} - \theta) = \left(-\frac{1}{n} \sum_{i \in \mathcal{M}} \frac{\partial H_{i\theta}}{\partial \theta^{\mathrm{T}}} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} H_{i\theta} + o_p(1). \tag{13}$$

Let

$$\begin{aligned} A_{int} &= \lim_{n \to \infty} E \left(-\frac{1}{n} \sum_{i \in \mathcal{M}} \frac{\partial H_{i\theta}}{\partial \theta^T} \right), \\ B_{int} &= \lim_{n \to \infty} \operatorname{var} \left(\frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} H_{i\theta} \right). \end{aligned}$$

Noting that $\partial S_{\mathrm{V}i\beta} / \partial \alpha^{\mathrm{T}} = 0$ and $\partial S_{\mathrm{V}i\alpha} / \partial \beta^{\mathrm{T}} = 0$, we obtain A_{int} in (14):

$$\begin{aligned} A_{int} &= \lim_{n \to \infty} E \left\{ -\frac{1}{n} \left(\begin{array}{cc} \sum_{i \in \mathcal{M} \backslash \mathcal{V}} \frac{\partial S_{\mathrm{M}i\beta}}{\partial \beta^{\mathrm{T}}} + \sum_{i \in \mathcal{V}} \frac{\partial S_{\mathrm{V}i\beta}}{\partial \beta^{\mathrm{T}}} & \sum_{i \in \mathcal{M} \backslash \mathcal{V}} \frac{\partial S_{\mathrm{M}i\beta}}{\partial \alpha^{\mathrm{T}}} \\ \sum_{i \in \mathcal{M} \backslash \mathcal{V}} \frac{\partial S_{\mathrm{M}i\alpha}}{\partial \beta^{\mathrm{T}}} & \sum_{i \in \mathcal{M} \backslash \mathcal{V}} \frac{\partial S_{\mathrm{M}i\alpha}}{\partial \alpha^{\mathrm{T}}} + \sum_{i \in \mathcal{V}} \frac{\partial S_{\mathrm{V}i\alpha}}{\partial \alpha^{\mathrm{T}}} \end{array} \right) \right\} \\ &= -(1 - \rho) E \left(\frac{\partial S_{\mathrm{M}i\theta}}{\partial \theta^{\mathrm{T}}} \right) - \rho \left(\begin{array}{cc} E \left(\frac{\partial S_{\mathrm{V}i\beta}}{\partial \beta^{\mathrm{T}}} \right) & 0 \\ 0 & E \left(\frac{\partial S_{\mathrm{V}i\alpha}}{\partial \alpha^{\mathrm{T}}} \right) \end{array} \right), \end{aligned} \tag{14}$$

and

$$\begin{aligned}
B_{int} &= \lim_{n\to\infty}\frac{1}{n}\sum_{i\in\mathcal{M}}E\begin{pmatrix} H_{i\beta}H_{i\beta}^{\mathrm{T}} & H_{i\beta}H_{i\alpha}^{\mathrm{T}} \\ H_{i\alpha}H_{i\beta}^{\mathrm{T}} & H_{i\alpha}H_{i\alpha}^{\mathrm{T}} \end{pmatrix} \\
&= \lim_{n\to\infty}\frac{n-n_{\mathrm{V}}}{n}\frac{1}{n-n_{\mathrm{V}}}\sum_{i\in\mathcal{M}\backslash\mathcal{V}} \\
&\quad E\begin{pmatrix} S_{\mathrm{M}i\beta}S_{\mathrm{M}i\beta}^{\mathrm{T}} & S_{\mathrm{M}i\beta}S_{\mathrm{M}i\alpha}^{\mathrm{T}} \\ S_{\mathrm{M}i\beta}S_{\mathrm{M}i\beta}^{\mathrm{T}} & S_{\mathrm{M}i\alpha}S_{\mathrm{M}i\alpha}^{\mathrm{T}} \end{pmatrix} \\
&\quad + \lim_{n\to\infty}\frac{n_{\mathrm{V}}}{n}\cdot\frac{1}{n_{\mathrm{V}}}\sum_{i\in\mathcal{V}}E\begin{pmatrix} S_{\mathrm{V}i\beta}S_{\mathrm{V}i\beta}^{\mathrm{T}} & 0 \\ 0 & S_{\mathrm{V}i\alpha}S_{\mathrm{V}i\alpha}^{\mathrm{T}} \end{pmatrix} \\
&= (1-\rho)E(S_{\mathrm{M}i\theta}S_{\mathrm{M}i\theta}^{\mathrm{T}}) \\
&\quad +\rho\begin{pmatrix} E(S_{\mathrm{V}i\beta}S_{\mathrm{V}i\beta}^{\mathrm{T}}) & 0 \\ 0 & E(S_{\mathrm{V}i\alpha}S_{\mathrm{V}i\alpha}^{\mathrm{T}}) \end{pmatrix}
\end{aligned}$$

Since $A_{int} = B_{int}$, then applying the Central Limit Theorem to (13) gives that

$$\sqrt{n}(\hat{\theta}_{int}-\theta) \xrightarrow{d} N(0, A_{int}^{-1}) \quad \text{as} \quad n\to\infty.$$

APPENDIX C

Under regularity conditions, maximizing the likelihood with respect to the parameters is equivalent to solving the equation

$$\sum_{i\in\mathcal{M}\cup\mathcal{V}}\begin{pmatrix} (1-\delta_i)S_{\mathrm{M}i\beta} \\ (1-\delta_i)S_{\mathrm{M}i\alpha}+\delta_i S_{\mathrm{V}i\alpha}^{*} \end{pmatrix} = 0.$$

Let $H_{i\beta} = (1-\delta_i)S_{\mathrm{M}i\beta}$, $H_{i\alpha} = (1-\delta_i)S_{\mathrm{M}i\alpha}+\delta_i S_{\mathrm{V}i\alpha}^{*}$ and $H_{i\theta} = (H_{i\beta}^{\mathrm{T}}, H_{i\alpha}^{*\mathrm{T}})^{\mathrm{T}}$. Equation (13) now becomes

$$\sqrt{n+n_{\mathrm{V}}}(\hat{\theta}_{ext}-\theta) =$$
$$\left(-\frac{1}{n+n_{\mathrm{V}}}\sum_{i\in\mathcal{M}\cup\mathcal{V}}\frac{\partial H_{i\theta}}{\partial\theta^{\mathrm{T}}}\right)^{-1}\frac{1}{\sqrt{n+n_{\mathrm{V}}}}\sum_{i\in\mathcal{M}\cup\mathcal{V}}H_{i\theta}+o_p(1).$$

Rewrite this as

(15)
$$\sqrt{\frac{n_{\mathrm{V}}}{n}+1}\sqrt{n}(\hat{\theta}_{ext}-\theta) =$$
$$\left(-\frac{1}{n+n_{\mathrm{V}}}\sum_{i\in\mathcal{M}\cup\mathcal{V}}\frac{\partial H_{i\theta}}{\partial\theta^{\mathrm{T}}}\right)^{-1}\frac{1}{\sqrt{n+n_{\mathrm{V}}}}\sum_{i\in\mathcal{M}\cup\mathcal{V}}H_{i\theta}+o_p(1),$$

and let

$$A_{ext} = \lim_{n\to\infty}E\left(-\frac{1}{n+n_{\mathrm{V}}}\sum_{i\in\mathcal{M}\cup\mathcal{V}}\frac{\partial H_{i\theta}}{\partial\theta^{T}}\right),$$

and

$$B_{ext} = \lim_{n\to\infty}\mathrm{var}\left(\frac{1}{\sqrt{n+n_{\mathrm{V}}}}\sum_{i\in\mathcal{M}\cup\mathcal{V}}H_{i\theta}\right).$$

Then similar calculations to A_{int} and B_{int} give

$$A_{ext} = -\frac{1}{(1+\rho)}E\left(\frac{\partial S_{\mathrm{M}i\theta}}{\partial\theta^{\mathrm{T}}}\right) - \frac{\rho}{1+\rho}\begin{pmatrix} 0 & 0 \\ 0 & E\left(\frac{\partial S_{\mathrm{V}i\alpha}^{*}}{\partial\alpha^{\mathrm{T}}}\right) \end{pmatrix},$$

and

$$B_{ext} = \frac{1}{(1+\rho)}E(S_{\mathrm{M}i\theta}S_{\mathrm{M}i\theta}^{\mathrm{T}}) + \frac{\rho}{1+\rho}\begin{pmatrix} 0 & 0 \\ 0 & E(S_{\mathrm{V}i\alpha}^{*}S_{\mathrm{V}i\alpha}^{*\mathrm{T}}) \end{pmatrix}.$$

Since $A_{ext} = B_{ext}$, then applying the Central Limit Theorem to (15) gives that

$$\sqrt{n}(\hat{\theta}_{ext}-\theta) \xrightarrow{d} N\left(0, \frac{1}{1+\rho}A_{ext}^{-1}\right) \quad \text{as} \quad n\to\infty.$$

ACKNOWLEDGEMENTS

The authors thank the Associate Editor and the referees for their comments on the initial version. The research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Received 14 September 2017

REFERENCES

Andersen, P. K., Borgan, Ø, Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag New York, Inc. MR1198884

Bross, I. (1954). Misclassification in 2 × 2 tables. *Biometrics*, 10(4):478–486. MR0068796

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications.* Chapman & Hall/CRC. MR2682774

Carroll, R. J., Gail, M. H., and Lubin, J. H. (1993). Case–control studies with errors in covariates. *Journal of the American Statistical Association*, 88, 185–199.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceany, C. M. (2006). *Measurement Error in Nonlinear Models*, 2nd ed., Chapman & Hall.

Chen, B., Yi, G. Y., and Cook, R. J. (2010). Analysis of interval-censored disease progression data via multi-state models under a nonignorable inspection process. *Statistics in Medicine*, 29(11):1175–1189.

Chen, P.-L. and Sen, P. K. (1999). A piecewise transition model for analyzing multistate life history data. *Journal of Statistical Planning and Inference*, 78(1–2):385–400. MR1705558

Chiang, C.-L. (1980). *An Introduction to Stochastic Processes and Their Application.* Huntington, N.Y.: Krieger.

Cook, R. J., Kalbfleisch, J., D., and Yi, G. Y. (2002). A generalized mover-stayer model for panel data. *Biostatistics*, 3(3):407–420.

Cook, R. J., Zeng, L., and Lee, K.-A. (2008). A multistate model for bivariate interval censored failure time data. *Biometrics*, 64(4):1100–1109.

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes.* London: Methuen.

Espeland, M. A. and Hui, S. L. (1987). A general approach to analyzing epidemiologic data that contain misclassification errors. *Biometrics*, 43(4):1001–1012.

Gentleman, R. C., Lawless, J. F., Lindsey, J. C., and Yan, P. (1994). Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine*, 13(8):805–821.

Greenland, S. (1988). Statistical uncertainty due to misclassitication: Implications for validation substudies. *Clinical Epidemiology*, 41, 1167–1174.

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology*. Chapman & Hall/CRC, Boca Raton, Florida.

Hsieh, H.-J., Chen, T. H.-H., and Chang, S.-H. (2002). Assessing chronic disease progression using non-homogeneous exponential regression Markov models: an illustration using a selective breast cancer screening in Taiwan. *Statistics in Medicine*, 21(22):3369–3382.

Hubbard, R. A., Inoue, L. Y. T., and Fann, J. R. (2008). Modeling nonhomogeneous Markov processes via time transformation. *Biometrics*, 64(3):843–850. MR2526635

Jackson, C. H. (2011). Multi-state models for panel data: the msm package for R. *Journal of Statistical Software*, 38(8):1–28.

Kalbfleisch, J. D. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80(392):863–871.

Kosorok, M. R. and Chao, W.-H. (1996). The analysis of longitudinal ordinal response data in continuous time. *Journal of the American Statistical Association*, 91(434):807–817.

Küchenhoff, H., Lederer, W., and Lesaffre, E. (2007). Asymptotic variance estimation for the misclassification simex. *Computational Statistics & Data Analysis*, 51(12):6197–6211.

Küchenhoff, H., Mwalili, S. M., and Lesaffre, E. (2006). A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics*, 62(1):85–96.

Kuha, J., Skinner, C., and Palmgren, J. (2005). Misclassification error. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*. Chichester, West Sussex, England; Hoboken, NJ: Wiley, 2nd edition.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. Spinger-Verlag, New York, LLC.

Lindsey, J. C. and Ryan, L. M. (1993). A three-state multiplicative model for rodent tumorigenicity experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 42(2):283–300.

Marshall, R. J. (1990). Validation study methods for estimating exposure proportions and odds ratios with misclassified data. *Journal of Clinical Epidemiology*, 43(9):941–947.

Morrissey, M. J. and Spiegelman, D. (1999). Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. *Biometrics*, 55(2):338–344.

Rosychuk, R. J. and Thompson, M. E. (2003). Bias correction of two-state latent Markov process parameter estimates under misclassification. *Statistics in Medicine*, 22, 2035–2055.

Rosychuk, R. J. and Thompson, M. E. (2004). Parameter identifiability issues in a latent Markov model for misclassified binary responses. *Journal of Iranian Statistical Society*, 3, 39–57.

Saint-Pierre, P., Combescure, C., Daurès, J. P., and Godard, P. (2003). The analysis of asthma control under a Markov assumption with use of covariates. *Statistics in Medicine*, 22, 3755–3770.

Spiegelman, D., Carroll, R. J., and Kipnis, V. (2001). Efficient regression calibration for logistic regression in main study/internal validation study designs. *Statistics in Medicine*, 29, 139–160.

Spiegelman, D., Rosner, B., and Logan, R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association*, 95, 51–61.

Tom, B. D. M. and Farewell, V. T. (2011). Intermittent observation of time-dependent explanatory variables: a multistate modelling approach. *Statistics in Medicine*, 30, 3520–3531.

van den Hout, A. and Matthews, F. E. (2008). Multi-state analysis of cognitive ability data: A piecewise-constant model and a Weibull model. *Statistics in Medicine*, 27, 5440–5455.

Willett, W. C. (1998). *Nutritional Epidemiology*. 2nd ed. Oxford University Press: New York.

Yi, G. Y. (2017). *Statistical Analysis with Measurement Error or Misclassication: Strategy, Method and Application.*. Springer Science+Business Media LLC, New York.

Yi, G. Y., He, W., and He, F. (2017). Analysis of panel data under hidden mover-stayer models. *Statistics in Medicine*, 36, 3231–3243.

Yi, G. Y., Ma, Y., Spiegelman, D., and Carroll, R. J. (2015). Functional and structural methods with mixed measurement error and misclassication in covariates. *Journal of the American Statistical Association*, 110, 681–696.

Yi, G. Y., Yan, Y., Liao, X., and Spiegelman, D. (2019). Parametric Regression Analysis with Covariate Misclassification in Main Study/Validation Study Designs. To appear in *The International Journal of Biostatistics*.

Grace Yi
Department of Statistics and Actuarial Science
University of Waterloo
200 University Avenue West
Waterloo, Ontario
Canada, N2L 3G1
E-mail address: yyi@uwaterloo.ca

Wenqing He
Department of Statistical and Actuarial Sciences
University of Western Ontario
1151 Richmond Street North
London, Ontario
Canada, N6A 5B7
E-mail address: whe@stats.uwo.ca

Feng He
Department of Statistics and Actuarial Science
University of Waterloo
200 University Avenue West
Waterloo, Ontario
Canada, N2L 3G1
E-mail address: feng.he.stat@gmail.com