

# A hidden Markov model approach to analyze longitudinal ternary outcomes when some observed states are possibly misclassified

Julia S. Benoit,<sup>a,b</sup> Wenyaw Chan,<sup>b\*†</sup> Sheng Luo,<sup>b</sup>  
Hung-Wen Yeh<sup>c</sup> and Rachele Doody<sup>d</sup>

Understanding the dynamic disease process is vital in early detection, diagnosis, and measuring progression. Continuous-time Markov chain (CTMC) methods have been used to estimate state-change intensities but challenges arise when stages are potentially misclassified. We present an analytical likelihood approach where the hidden state is modeled as a three-state CTMC model allowing for some observed states to be possibly misclassified. Covariate effects of the hidden process and misclassification probabilities of the hidden state are estimated without information from a ‘gold standard’ as comparison. Parameter estimates are obtained using a modified expectation-maximization (EM) algorithm, and identifiability of CTMC estimation is addressed. Simulation studies and an application studying Alzheimer’s disease caregiver stress-levels are presented. The method was highly sensitive to detecting true misclassification and did not falsely identify error in the absence of misclassification. In conclusion, we have developed a robust longitudinal method for analyzing categorical outcome data when classification of disease severity stage is uncertain and the purpose is to study the process’ transition behavior without a gold standard. Copyright © 2016 John Wiley & Sons, Ltd.

**Keywords:** longitudinal data analysis; hidden Markov model; disease progression; misclassification

## 1. Introduction

Early disease detection is fundamental in improving medical treatment design and intervention aimed at delaying disease progression, subsequently enhancing quality of life and, in some cases, reducing disease mortality. Unfortunately, disease staging may be subject to misclassification because true events are not directly observable. Proxy variables help explain unobservable phenomena in medical research but introduce biased estimates and can be especially concerning when the misclassified observable outcome is categorical. Solutions for inaccurate continuous outcomes usually include a random effect or the like in longitudinal settings. Remedies for categorical outcomes potentially observed with error continue to be studied. Further challenging is the lack of a ‘gold standard’ for the targeted process (eg., Alzheimer’s disease (AD) staging). Multi-state transitional models are useful for quantifying disease staging and focus on the movement from one category where the interest lies in estimating the transition rates or intensities. Transition modeling approaches exist to examine misclassification in longitudinal studies where the outcome is categorical in both continuous and discrete-time settings,

<sup>a</sup>Texas Institute for Measurement, Evaluation, and Statistics and Department of Basic Vision Sciences, College of Optometry, The University of Houston, Houston, TX, 77204, U.S.A.

<sup>b</sup>Department of Biostatistics, The University of Texas Health Science Center at Houston, Houston, TX, 77030, U.S.A.

<sup>c</sup>Department of Biostatistics, The University of Kansas Medical Center, Kansas City, KS, 66160, U.S.A.

<sup>d</sup>Alzheimer’s Disease and Memory Disorders Center, Department of Neurology, Baylor College of Medicine, Houston, TX, 77030, U.S.A.

\*Correspondence to: Wenyaw Chan, Department of Biostatistics, The University of Texas Health Science Center at Houston, Houston, TX 77030, U.S.A.

†E-mail: wenyaw.chan@uth.tmc.edu

specifically hidden Markov models (HMMs). Discrete-time HMMs to examine misclassification have been considered by several authors ([1–5]). Focusing on continuous-time setting, two-state (binary outcome) Markov models accounting for misclassification have been studied using Bayesian [6,7] and classic expectation-maximization (EM) approaches [8] among others [9]. Multi-state HMMs (more than two states) are more complicated and, to date, a reversible HMM with an analytical solution to simultaneously estimate transition rates and probability of misclassification (or sensitivity) has not been developed. Special cases of the multi-state continuous-time Markov chain (CTMC) approaches have been developed including semi-Markov [10] and irreversible hidden CTMC models [11].

Because of the complexity of the likelihood function of a general multi-state recurrent CTMC, methodology for exact solutions when the target state is observable is in its infancy. Li and Chan [12] developed a likelihood technique to estimate the transition rates of a three-state CTMC with a binary covariate, thus providing comparisons of transition probabilities between states for two groups. The aforementioned likelihood technique was later extended to include multiple covariates, and also a practical interpretation of the process with covariates was provided [13,14]. This research focuses on a possibly misclassified ternary recurrent outcome observed at irregular and varying time intervals among each individual. We propose methodology that accounts for possible misclassification of the outcomes modeled as a CTMC and further generalize the Baum–Welch algorithm to the three-state continuous-time Markov model with covariates.

In circumstances where data are not directly observable, (e.g., misclassified outcomes), unique parameter estimates may be difficult or impossible to identify (i.e., non-identifiable) in the case of ‘blocked’ or ‘hidden’ information. Parameter non-identifiability has been discussed in the literature with regard to HMMs [8,15,16] among others, and is addressed in this research.

The three-state CTMC with one state subject to misclassification model, its likelihood, and the estimation method are described in the following section (Section 2). In Section 3, we describe the implementation of our method using a modified EM algorithm for parameter estimation and conduct a simulation study to assess its performance. Applications of our method to analyze AD caregiver stress-levels are described, and results are presented in Section 4. We make remarks on identifiability of CTMC estimation in Section 5. This paper concludes with a discussion of our findings in Section 6.

## 2. Methods

### 2.1. Ternary outcome with possible misclassification

Consider a longitudinal study where  $Z_k(t_{k,1}), Z_k(t_{k,2}), \dots, Z_k(t_{k,n_k})$  is a sequence of observed ternary outcomes recorded as 1, 2, or 3 and measured at times  $t_{k,1}, t_{k,2}, \dots, t_{k,n_k}$ , on subject  $k$ , and  $n_k$  is the number of observations on subject  $k$ . The observed sequence provides information for the ‘hidden’ sequence  $Y_k(t_{k,1}), Y_k(t_{k,2}), \dots, Y_k(t_{k,n_k})$  at time  $t_{k,1}, t_{k,2}, \dots, t_{k,n_k}$  for subject  $k$ , assumed to be measured with possible misclassification, which will be modeled as a three-state CTMC with state occupancy valued 1, 2, or 3. We assume dependency of the observed state on the state of the ‘hidden’ process solely at matching time points, not on the previous history of either the observed or hidden processes formulated as follows:

$$\begin{aligned} & \Pr(Z_k(t_{k,s})|Y_k(t_{k,1}), \dots, Y_k(t_{k,s}), Z_k(t_{k,1}), \dots, Z_k(t_{k,s-1})) \\ &= \Pr(Z_k(t_{k,s})|Y_k(t_{k,s})) \\ &= \varepsilon_{Y(t_{k,s}), Z(t_{k,s})}, s = 1, \dots, n_k. \end{aligned} \tag{1}$$

Furthermore, Eq. 1 defines the probability that the observed state correctly classifies the hidden state of the process (or misclassification probability) {e.g.,  $\Pr(Z_k(t_{k,s})=j|Y_k(t_{k,s})=i)=\varepsilon_{ij}$  and when  $i=j$ ,  $\varepsilon_{ii}$  is the probability of correctly classifying (or identifying) the ‘hidden’ outcome.} Also, note

$$\text{that } \varepsilon_{ii} = 1 - \sum_{i=1, i \neq l}^3 \varepsilon_{il}.$$

The three-state CTMC is fully described by the instantaneous transition rates,  $q_{ij}$ , the rate at which the process transitions from state ‘ $i$ ’ to state ‘ $j$ ’, where  $i, j=1,2,3$ , and  $i \neq j$ . The infinitesimal matrix  $R$  is formed by these parameters:

$$R = \begin{bmatrix} -(q_{12} + q_{13}) & q_{12} & q_{13} \\ q_{21} & -(q_{21} + q_{23}) & q_{23} \\ q_{31} & q_{32} & -(q_{31} + q_{32}) \end{bmatrix} \quad (2)$$

From the property of a CTMC, the sojourn time or amount of time a process stays in category ‘*i*’ before exiting follows an exponential distribution with mean  $\left(\sum_{l=1, l \neq i}^3 q_{il}\right)^{-1}$  and is generally unobservable. At transition time, the probability of transitioning into state ‘*j*’ given that the process is currently in state ‘*i*’ is calculated as  $Q_{ij} = q_{ij} \left(\sum_{l=1, l \neq i}^3 q_{il}\right)^{-1}$ . The transition rates make up the probability mechanism used to derive the probability of transition over a specific interval of time,  $P_{ij}(t)$ . Explicit algebraic formulas were derived for three scenarios [12] allowing us to write the likelihood function in terms of the transition probabilities, which are functions of the  $q_{ij}$  parameters to be estimated.

The distribution of the initial observations, denoted as  $\pi_i = P\{Y(0) = i\}$ , where ‘*i*’ takes on the values 1, 2, or 3, under our model are also assumed indirectly observable as in Eq. 1. We transformed the transition rates via a log-link function to examine the linear combination of the covariates  $x_r$  by estimating coefficients  $\beta_r$  and each unique intercept parameter  $\alpha_{ij}$  for each  $q_{ij}$ , that is,

$$\log q_{ij} = \alpha_{ij} + \sum_{r=1}^p \beta_r x_r, \text{ for } i \neq j, i, j = 1, 2, 3, \text{ and } r = 1, \dots, p. \quad (3)$$

By the Markov property of the hidden process and the basic probability, we can construct the likelihood function of  $Z_k(t_{k,1}), Z_k(t_{k,2}), \dots, Z_k(t_{k,n_k})$  as

$$\begin{aligned} &P(Z(t_1) = z(t_1), Z(t_2) = z(t_2), \dots, Z(t_{n_k}) = z(t_{n_k})) \\ &= \sum_{\text{all possible } y\text{'s}} \varepsilon_{y(t_1)z(t_1)} \varepsilon_{y(t_2)z(t_2)} \dots \varepsilon_{y(t_{n_k})z(t_{n_k})} \times P(Y(t_1) = y(t_1)) \\ &\quad \times P_{y(t_1)y(t_2)}(t_2 - t_1) \dots P_{y(t_{n_k-1})y(t_{n_k})}(t_{n_k} - t_{n_k-1}), \end{aligned} \quad (4)$$

where  $\varepsilon_{y(t_1)z(t_1)} = 1$  and  $P_{y(t_m)y(t_{m+1})}(t_{m+1} - t_m)$  represents the likelihood that the hidden process is  $Y(t_{m+1})$  at time  $t_{m+1}$  given that  $Y(t_m)$  at time  $t_m$ . The complete likelihood function for all  $n_k$  observations for all subjects can be written as

$$L(\Theta) = \prod_{k=1}^m \left[ \sum_{\text{all } y\text{'s}} P(Y(t_{k,1}) = y(t_{k,1})) \left\{ \prod_{i=1}^{n_k} \left( \varepsilon_{Y(t_{k,i})Z(t_{k,i})} \right) \left\{ \prod_{j=1}^{n_k-1} \left( P_{Y(t_{k,j})Y(t_{k,j+1})}(t_{k,j+1} - t_{k,j}) \right) \right\} \right\} \right], \quad (5)$$

where  $t_{k,j} = 0$ , for  $j = 1$ ,  $y_k(t_{k,l})$ , denotes the category of the targeted outcome that would have been observed for the  $k^{th}$  individual at time  $t_{k,l}$ ,  $\pi_l = P(Y(t_{k,1}) = l)$ ,  $l = 1, 2, 3$  are the true initial distribution that will be treated as nuisance parameters, although their estimates will be obtained at each iteration in our estimation process for helping perform the EM algorithm. Note that  $\Theta = (\alpha_{12}, \alpha_{13}, \alpha_{21}, \alpha_{23}, \alpha_{31}, \alpha_{32}, \beta_1, \beta_2, \varepsilon_{21}, \varepsilon_{23})'$  is the vector of parameters defined prior to Eq. 5. Some components of  $\Theta$  are hidden in probability expression  $P_{Y(t_{k,j})Y(t_{k,j+1})}(t_{k,j+1} - t_{k,j})$ .

## 2.2. Estimation

The expectation of the complete log likelihood given the observed data and the parameter values of the  $v^{th}$  iteration can be expressed as

$$Q(\theta, \theta^{(v)}) = E[\log(L)|all\ z's, \theta^{(v)}] \\ = \sum_{k=1}^m E \left[ \log \left\{ \sum_{all\ y's} P(Y(t_{k,1}) = y(t_{k,1})) \left\{ \prod_{i=1}^{n_k} (\varepsilon_{Y(t_{k,i})Z(t_{k,i})}) \left\{ \prod_{j=1}^{n_k-1} (P_{Y(t_{k,j})Y(t_{k,j+1})}(t_{k,j+1}-t_{k,j})) \right\} \right\} \right\} | all\ z's, \theta^{(v)} \right] \quad (6)$$

For calculation reduction during maximization, a modified Baum–Welch [17] algorithm is proposed and derived beginning with the following forward-backward variables as

$$A_j(i) = P(Z(t_1) = z(t_1), Z(t_2) = z(t_2), \dots, Z(t_j) = z(t_j), Y(t_j) = i | \theta), \quad (7a)$$

$$B_j(i) = P(Z(t_{j+1}) = z(t_{j+1}), Z(t_{j+2}) = z(t_{j+2}), \dots, Z(t_{n_k}) = z(t_{n_k}) | \theta, Y(t_j) = i) \quad (7b)$$

Equation 7a expresses the forward variable as the probability of the partially observed sequence until time  $t_j$ , and the true state  $i$  at time  $t_j$  and Eq. 7b the backward variable as the probability of the partial observation sequence from  $t_{j+1}$  to the end, given the true state  $i$  at time  $t_j$  and the model, respectively. By routine algebra, conditional probabilities in Eq. 6 can be expressed as iteratively calculable terms:

$$P(Y(t_j) = w | all\ z's, \theta^{(v)}) = \frac{A_j(w)B_j(w)}{\sum_{l=1}^3 A_{n_k}(l)} \quad (8a)$$

$$P(Y(t_{k,j-1}) = w, Y(t_{k,j}) = l | all\ z's, \theta^{(v)}) = \frac{A_{k,j-1}(w)P_{wl}(t_j - t_{j-1})\varepsilon_{l,z(t_j)}B_{k,j}(l)}{\sum_{i=1}^3 A_{k,n_k}(i)}, \quad (8b)$$

where

$$A_1(i) = \pi_i \varepsilon_{i,z(t_1)} \quad A_{j+1}(l) = \left[ \sum_{w=1}^3 A_j(w)P_{wl}(t_{j+1} - t_j) \right] \varepsilon_{l,z(t_{j+1})}, \quad j = 1, \dots, n_k-1 \quad (9a)$$

$$B_j(w) = \sum_l P_{wl}(t_{j+1} - t_j) \varepsilon_{l,z(t_{j+1})} B_{j+1}(l), j = 1, \dots, n_k-1, \text{ when } j = n_k, B_{j+1}(l) = 1 \text{ for all } l \\ = 1, 2, 3. \quad (9b)$$

Note that Eqs. 8 and 9 provide similar explanation as that of the Baum–Welch algorithm. For example,  $A_1(i)$  represents the probability that the true initial state is  $i$  and the initial observed state is  $z(t_1)$  and  $A_{j+1}(i)$  represents the probability that the true state at  $t_{j+1}$  is  $i$  and all the observed states up to time  $t_{j+1}$ , and is expressed in  $A_j(\cdot)$ . Similarly,  $B_j(w)$  represents the probability that the true state at  $t_j$  is  $w$  and all the observed states from time  $t_{j+1}$ , and is expressed in  $B_j(\cdot)$ . Equations 9a and 9b compute numerical weights for the ‘hidden’ probability components used to update  $Q(\theta, \theta^{(v)})$  at each iteration, and  $\pi_i$  is substituted by  $\hat{\pi}_i$  in the implementation of this EM process.

### 3. Simulation study and numeric estimation procedures

To evaluate the performance of the proposed method, a simulation study is conducted where 1000 data sets were generated with  $N=1000$  individuals. Assuming that transition rates, thus transition time depend on each individual’s covariates, one binary and one continuous, a three-state recurrent CTMC was simulated for each individual in each replicate. We assume the CTMC state is observable only at integer times  $0, 1, \dots, 10$ , not at the actual transition times. Conditional on the hidden state of the process at time  $t_l$ , the observed outcomes follow a multinomial distribution with parameters  $\varepsilon_{21}$  and  $\varepsilon_{23}$  denoting the probabilities of misclassifying hidden state ‘2’ as observed state ‘1’ or ‘3’. We performed the simulation under two scenarios: (i) assuming misclassification is equally likely (i.e.,  $\varepsilon_{21} = \varepsilon_{23}$ ) and denoted as  $\varepsilon$ ; (ii)  $\varepsilon_{21} \neq \varepsilon_{23}$ . Thus, when  $\varepsilon_{21} = \varepsilon_{23} = \varepsilon$ , it should be clear that  $1 - 2\varepsilon$  is the probability of the observed

data correctly classifying the hidden state ‘2’. Probability mechanisms imposed and examined in this study were (i)  $\varepsilon_{21} = 2\%$ ,  $\varepsilon_{23} = 3\%$ ; (ii)  $\varepsilon_{21} = 0\%$ ,  $\varepsilon_{23} = 0\%$ ; (iii)  $\varepsilon = 1\%$ ; (iv)  $\varepsilon = 5\%$ ; and (v)  $\varepsilon = 10\%$  using the proposed HMM. Noteworthy is that the possibility of misclassification has been restricted to hidden state ‘2’ for two reasons: (i) to ensure our model is identifiable and (ii) to relate our proposed model to the real data setting that stage 2 is more likely to be misclassified than the other two states. The EM algorithm was used to update the likelihood for maximization at each iteration via implementation of the modified Baum–Welch algorithm. Additionally, data imposed with misclassification mechanism (iii)  $\varepsilon = 1\%$  were estimated naively for comparison. The choice of sample size and number of observations reflects approximately the number of subjects collected in the AD study cohort described in Section 4. The number of visits reached up to 14 for some individuals. True process parameter values were chosen as  $\alpha_{12} = 2.9$ ,  $\alpha_{13} = 2.7$ ,  $\alpha_{21} = 2.5$ ,  $\alpha_{23} = 2.2$ ,  $\alpha_{31} = 1.8$ ,  $\alpha_{32} = 1.1$ ,  $\beta_1 = 1$ ,  $\beta_2 = -0.5$  and calculated using a subset of data from the aforementioned dataset with disease staging as the outcome.

To improve the computational efficiency, we derived a data-based procedure for calculating initial parameter values for the EM estimation. All possible combinations of individual covariates and their transition rate relationship from Eq. 3 provided maximum likelihood estimates (MLEs) for  $q_{ij}$  and were then post-estimated to obtain regression coefficients for each parameter. By the same notion, starting values for misclassification parameters were obtained via post-estimation following our proposed polytomous logistic regression with covariates to predict the true state, and thus, the proportion of misclassification from true state 2 was used as the starting parameter. Quasi-Newton optimization was used to find the MLE of the parameter set of the log-likelihood function and then compared with the true

**Table I.** Simulation results of proposed and naive methods with 1% imposed misclassification under equally likely scenario (i.e.,  $\varepsilon_{21} = \varepsilon_{23} = \varepsilon = 1\%$ ).

Parameter	True	Proposed method**				Naive approach**			
		Estimate	% Bias	CP (%)	Avg SE	Estimate	% Bias	CP (%)	Avg SE
$\alpha_{12}$	2.9	2.91	0.45	93.9	0.18	2.45	-15.5	21.2	0.16
$\alpha_{13}$	2.7	2.71	0.44	94.4	0.16	2.23	-17.6	12.9	0.15
$\alpha_{21}$	2.5	2.51	0.48	94.3	0.18	2.10	-16.0	29.9	0.16
$\alpha_{23}$	2.2	2.21	0.44	94.0	0.16	1.88	-14.5	31.6	0.14
$\alpha_{31}$	1.8	1.81	0.76	94.5	0.16	1.31	-27.2	10.0	0.15
$\alpha_{32}$	1.1	1.10	-0.21	92.9	0.18	0.84	-23.3	52.5	0.16
$\beta_1$	1.00	0.998	-0.16	93.5	0.08	0.89	-11.0	55.4	2.4
$\beta_2$	-0.50	-0.50	0.29	93.9	0.02	-0.44	-12.4	3.4	0.01
$\varepsilon$	0.01	0.01	-0.22	73.9	0.001				

\*1% misclassification refers to the equally likely probability of target state ‘2’ misclassified as ‘1’ or ‘3’ and 1-2  $\varepsilon$  is the probability of target state ‘2’ correctly observed as ‘2’.

\*\*Hidden method yielded 99.5% estimable datasets; naive yielded 76% estimable datasets.

SE-standard error; CP-coverage probability.

**Table II.** Simulation results of proposed methods with 5% and 10% imposed misclassification under equally likely scenario (i.e.,  $\varepsilon_{21} = \varepsilon_{23} = \varepsilon = 5\%$  and 10%).

Parameter	True	Misclassification rate = 5%				Misclassification rate = 10%			
		Estimate	% Bias	CP (%)	Avg. SE	Estimate	% Bias	CP (%)	Avg. SE
$\alpha_{12}$	2.9	2.91	0.42	91.3	0.18	2.92	-0.62	90.0	0.18
$\alpha_{13}$	2.7	2.71	0.26	92.6	0.16	2.70	0.08	90.8	0.16
$\alpha_{21}$	2.5	2.51	0.40	91.3	0.18	2.51	0.46	91.1	0.18
$\alpha_{23}$	2.2	2.21	0.48	91.4	0.16	2.21	0.25	88.8	0.16
$\alpha_{31}$	1.8	1.81	0.63	92.0	0.16	1.81	0.54	89.9	0.16
$\alpha_{32}$	1.1	1.10	-0.44	92.3	0.18	1.08	-1.6	89.7	0.18
$\beta_1$	1.00	0.996	-0.40	92.8	0.08	0.99	-0.56	91.2	0.08
$\beta_2$	-0.50	-0.50	0.21	91.8	0.02	-0.50	0.02	89.8	0.02
$\varepsilon_2$	0.05/1	0.05	0.22	78.5	0.03	0.10	-0.33	77.8	0.004

SE-standard error; CP-coverage probability.

**Table III.** Simulation results of proposed methods with 5% imposed misclassification under  $\varepsilon_{21} + \varepsilon_{23} = 5\%$  scenario ( $\varepsilon_{21} = 2\%$ ,  $\varepsilon_{23} = 3\%$ ).

Parameter	True	Estimate	% Bias	CP (%)	Avg. SE
$\alpha_{12}$	2.9	2.91	0.48	92.4	0.18
$\alpha_{13}$	2.7	2.71	0.34	93.9	0.16
$\alpha_{21}$	2.5	2.51	0.47	93.5	0.18
$\alpha_{23}$	2.2	2.21	0.41	93.4	0.16
$\alpha_{31}$	1.8	1.81	0.69	94.4	0.16
$\alpha_{32}$	1.1	1.10	-0.39	93.2	0.18
$\beta_1$	1.00	1.00	-0.15	93.9	0.08
$\beta_2$	-0.50	-0.50	0.26	93.9	0.02
$\varepsilon_{21}$	0.02	0.02	-0.21	73.3	0.002
$\varepsilon_{23}$	0.03	0.03	-0.07	78.2	0.003

SE-standard error; CP-coverage probability.

**Table IV.** Simulation results of proposed methods with 0% imposed misclassification under  $\varepsilon_{21} + \varepsilon_{23} = 0\%$  scenario ( $\varepsilon_{21} = 0\%$ ,  $\varepsilon_{23} = 0\%$ ).

Parameter	True	Estimate	% Bias	CP (%)	Average SE
$\alpha_{12}$	2.9	2.92	0.59	95.0	0.18
$\alpha_{13}$	2.7	2.72	0.57	94.7	0.16
$\alpha_{21}$	2.5	2.51	0.55	94.9	0.18
$\alpha_{23}$	2.2	2.21	0.63	94.7	0.16
$\alpha_{31}$	1.8	1.82	1.03	94.9	0.16
$\alpha_{32}$	1.1	1.10	0.00	95.0	0.18
$\beta_1$	1.00	1.00	-0.04	95.0	0.08
$\beta_2$	-0.50	-0.50	0.40	94.9	0.02
$\varepsilon_{21}$	0	0.0002	*	94.1	0.0002
$\varepsilon_{23}$	0	0.0001	*	98.5	0.0001

\*True parameter is '0', thus incalculable.

SE-standard error; CP-coverage probability.

parameters using bias and coverage probability measures. For each parameter, empirical means, biases, standard errors, and coverage probabilities for estimation are presented in Tables I–IV.

For 1000 replicates, Tables I–III show the bias of all estimated coefficients are negligible, and the coverage probabilities are mostly above 90% for our proposed method under both misclassification scenarios (i.e.,  $\varepsilon_{21} = \varepsilon_{23}$  and  $\varepsilon_{21} \neq \varepsilon_{23}$ ) and of varying rates (1%, 5%, and 10%). The biases of the misclassification parameter estimates are small, and the coverage probabilities range from roughly 75% to 79%. Additionally, coverage probabilities of the estimated misclassification parameter increase with the rate of misclassification (Tables I–II) suggesting that our method can detect sizeable amounts of misclassified outcomes; however, as the misclassification rate increases, it may be difficult to capture the true parameters. The naive analysis (Table I) with only a 1% imposed error rate yielded larger bias, poor coverage probability (3–52%), and only 77% estimable data replicates, confirming that when data truly are misclassified using the naive approach could lead to inaccurate conclusions. The misclassification modeling scenario where  $\varepsilon_{21} \neq \varepsilon_{23}$  performs about the same as when  $\varepsilon_{21} = \varepsilon_{23}$  (Tables II–III) with respect to both estimated coefficients and misclassification rate estimates. Finally, Table IV demonstrates that when categorical outcomes are classified correctly, our proposed method does not falsely provide a significant misclassification rate (a.k.a. specificity).

We have also conducted (results not presented) simulation studies to test the robustness of heterogeneities in progression rates. Specifically, we switched the direction of the process moving from state 3 to have a higher weight of moving to '2' instead of '1'. Additionally, we have conducted a simulation study to assess the method with a small sample size. The results are similar. That means the proposed model is robust to at least some changes of parameters.

#### 4. Longitudinal Alzheimer's disease study caregiver-stress levels

The proposed method's ability to assess sensitivity of longitudinal outcome data without a 'gold standard' is presented here. Data collected from January 1990 to September 2011 were extracted from the Baylor

Alzheimer’s Disease and Memory Disorders Center. Patients referred and self-referred to the center with probable AD, defined using criteria from the National Institute of Neurological and Communicative Disorders and Stroke [18], were used in this study. Patients underwent comprehensive evaluation and socio-demographic information such as age, sex and years of education, medical history, and estimates of symptom duration [19] were collected at baseline. Further details regarding the baseline assessment, follow-up, and outcome diagnosis have been described elsewhere [20]. Intervals of time between visits varied among individuals as did the number of visits themselves. Patients were neuropsychologically evaluated at baseline and annually or on an as-needed basis for medication management. The mini-mental state examination (MMSE) [21] was among the tests implemented and aids in identifying dementia progression and severity, focusing on memory, attention, and language. Scores range from 0 to 30 with lower scores indicating severe dementia. Additionally, the caregiver, a family member or friend spending the most time with the patient, provided information on their health and well-being. In this study, longitudinal self-rated stress levels of each caregiver were modeled as a CTMC with three categories (mild, moderate, and severe). Covariates baseline MMSE and caregiver relationship to the patient (spouse vs. other) were examined to better understand the movement between stages of AD patient caregiver stress-levels. Self-reported stress levels were based on a construct of a caregiver questionnaire. Patients with complete information on baseline MMSE, relationship to the caregiver, and whose caregivers provided at least two self-rated stress levels (i.e., at least one possible transition) were included in this analysis. Inter-observation time was calculated as the duration between two consecutive observations. This model’s ability to measure transitions over time while accounting for uneven intervals and number of observations allows the inclusion of patients with intermittent and monotone missing data under the assumption of missing completely at random.

A total of 952 patients had at least two caregiver self-rated stress levels and were included in the analysis. The average age of the patient was 74, ranging from 44–93, and the majority (68%) was female. The median number of visits was 3, ranging from 2–14, and baseline stress levels were distributed as 33% mild, 46% moderate, and 21% severe levels of stress. Three analyses were conducted to reflect the three models presented in Section 3. Note that we constrained  $q_{13}, q_{31} = 0$  because of the relative few transitions taking place from 1 (‘mild’) to 3 (‘severe’) and from 3 to 1 between two visits.

Table V displays the parameter estimates obtained from the two proposed methods and from the naive method that ignores the possibility of incorrectly observing 1 or 3. The main finding in this analysis is the significance of the misclassification parameters and the variations of the intensity parameter estimates between the proposed and naive methods. The overall estimate of correctly classified data differs between proposed methods M1 and M2 (M1: 1–0.05–0.10; M2: 1–2\*0.05), by about 5% (85% and 90%, respectively). When using all three approaches to estimate the movement between stages of AD patient caregiver stress-levels, parameters reflecting movement from state ‘1’ were not significant. After adjusting for patient relationship and baseline MMSE score, the proposed method suggests that at the time of change, a caregiver who is moderately stressed is more likely to reach to increase to a ‘severe’ stress level than revert to mildly stressed.

**Table V.** Comparison of proposed and naive approaches for estimating the parameters of movement between stages of Alzheimer’s disease patient caregiver stress levels.

Parameter	Proposed method M1			Proposed method M2			Naive		
	Estimate	SE	$e^{\alpha_{ij}}/e^{-\beta_r}$	Estimate	SE	$e^{\alpha_{ij}}/e^{-\beta_r}$	Estimate	SE	$e^{\alpha_{ij}}/e^{-\beta_r}$
$\alpha_{12}$	−0.43	0.37	0.65	−0.36	0.30	0.70	−0.23	0.19	0.79
$\alpha_{21}$	−3.48*	0.34	0.03	−3.63*	0.31	0.03	−0.77*	0.19	0.46
$\alpha_{23}$	−2.43*	0.29	0.09	−1.97*	0.25	0.14	−0.91*	0.19	0.40
$\alpha_{32}$	−1.10*	0.30	0.33	−0.50*	0.24	0.61	−0.04	0.18	0.96
$\beta_1$	−0.86*	0.17	2.36	−0.68*	0.14	1.97	−0.28*	0.10	1.32
$\beta_2$	−0.01	0.01	1.01	−0.00	0.01	1.00	−0.01	0.01	1.01
$\varepsilon_{21}$	0.05*	0.00	—	—	—	—	—	—	—
$\varepsilon_{23}$	0.10*	0.01	—	—	—	—	—	—	—
$\varepsilon_2$	—	—	—	0.05*	0.004	—	—	—	—

\*p-value <.05;

\*\*Note that  $e^{-\beta_r} = 1/e^{\beta_r}$  is the multiplicative impact on mean duration.

SE-standard error.

M1-scenario i; M2-scenario ii.

Note that the probability of this type of transition is  $0.74 \{ \exp(-2.43) / [\exp(-2.43) + \exp(-3.48)] \}$  and 0.47 for the naive approach. It is noteworthy to point out that the frequency of transitions from 'mild' to 'severe' and 'severe' to 'mild' is 1.8% and 2.4%, respectively, which is very few. The proposed method M1 suggests that caregivers accurately report their stress levels 85% of the time, and those patients moderately stressed are only slightly less likely to over-report their stress rather than under-report it.

## 5. Identifiability

Because of the nature of the process, a CTMC model may be non-identifiable when outcomes are recorded at pre-specified times with or without misclassification. If the sojourn time and state of change are recorded, the model will be fully identifiable. If the mean sojourn time is much longer than the inter-observation interval, the non-identifiable problem will be almost negligible. A more likely scenario is that the mean sojourn time interval ( $1/q_{ii}$ ) is shorter than the inter-observation time interval, and stage changes could be missed between observational periods. In other words, an observed transition (outcome) could be reached by at least two different paths. The complexity of this problem increases when a misclassification parameter is added to the model. Under certain conditions, the model specified in this paper can achieve a working level of identifiability. First, when state 2 is observed, it is not misclassified. So, an observed 2-to-1 transition can be used to compare with a 2-to-2 transition. If the true transitioned state is 2, the dynamic behavior of the observed 2-to-1 transition should be similar to that of the 2-to-2 transition. A similar comparison can also be applied to the observed 2-to-3 transition. Second, in our model, covariates  $X_1$  and  $X_2$  are not misclassified and are used to link the transition rates of the hidden states via a regression model. Thus, covariates can help amplify the inter-transition time and assist in identifying misclassification status. In other words, they can improve parameter identifiability. However, if all three states are possibly misclassified, this effect may be eliminated.

## 6. Discussion

We have proposed a method to estimate parameters of a hidden three-state CTMC with covariates when some observed outcomes are potentially misclassified and to estimate the probabilities of misclassification (sensitivity rates) in the absence of 'gold standard' information. We allow movement between all possible states. This estimation method was able to recover the hidden parameters with varying levels of misclassification imposed (up to a total of 20%). Simulation studies revealed the disruption of naively analyzing even a small amount of misclassified data with uneven observational schedules among and within patients (Table I). Another attractive feature of the proposed method is that it detected almost negligible error rates when the data were correctly classified and does not falsely provide a significant misclassification rate (a.k.a. specificity). Finally, we have shown, without presenting the results, that this method is robust to heterogeneities in the progression rates to smaller sample sizes ( $N=350$ ) and shorter chains (reduced to six observation times).

The complexity of statistical methodology with potentially misclassified outcomes and the drawbacks of ignoring this scenario are well known. For irreversible multi-state processes, approaches exist to account for potential misclassification, but, to date, no other methods take into account the more general case of recurrent multi-state processes subject to misclassification. Our analytical likelihood approach has been developed to confront this. The impact of this research contributes substantially to the situation where the outcome is truly hidden, and the validity of the observed data is suspect yet also provides a flexible approach when the researcher is unsure of the nature of whether the true state is misclassified or not.

We have shown that our proposed method gives very different results from the naive method in terms of significance and interpretation of covariate effects when using the naive approach to estimate the movement between stages of AD patient caregiver stress-levels.

One of the major difficulties in analyzing longitudinal categorical data using a Markov model is to mathematically prove the future distribution of the data depends on the present not the past at any time point. In our model, although the hidden process is assumed to be a CTMC, the observed data does not constitute a CTMC, and hence, it is not possible to validate the Markov property of the underlying process.

Finally, this method can be applied to any longitudinal discrete data with or without possible misclassification if the purpose is to look at transition behavior. Diseases are often referred to in terms of progression (irreversible disease processes) and is a special case of our method. Thus, our method can handle both general and special cases of the three-state CTMC with some observed states potentially misclassified.

## Acknowledgements

Julia Benoit and Wenyaw Chan were supported by NIH grant 2T32GM074902.

## References

1. Yeh H, Chan W, Symanski E. Intermittent missing observations in discrete-time hidden Markov models. *Communications in Statistics-Simulation and Computation* 2012; **41**:167–181.
2. Poskitt DS, Zhang J. Estimating components in finite mixtures and hidden Markov models. *Australian & New Zealand Journal of Statistics* 2005; **47**(3):269–286.
3. Altman RM, Petkau AJ. Applications of hidden Markov models to multiple sclerosis lesion count data. *Statistics in Medicine* 2005; **24**:2335–2344.
4. Le Strat Y, Carrat F. Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in Medicine* 1999; **18**:3463–3478.
5. Pfeffemann D, Skinner C, Humphreys K. The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 1998; **161**(1):13–32.
6. Smith T, Vounatsou P. Estimation of infection and recovery rates for highly polymorphic parasites when detectability is imperfect, using hidden Markov models. *Statistics in Medicine* 2003; **22**:1709–1724.
7. Rosychuk RJ, Thompson ME. A semi-Markov model for binary longitudinal responses subject to misclassification. *The Canadian Journal of Statistics* 2001; **29**(3):395–404.
8. Bureau A, Shiboski S, Hughes JP. Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Statistics in Medicine* 2003; **22**:441–462.
9. Rosychuk RJ, Sheng X, Stuber JL. Comparison of variance estimation approaches in a two-state Markov model for longitudinal data with misclassification. *Statistics in Medicine* 2006; **25**:1906–1921.
10. Van den Hout A, Matthews FE. Multi-state analysis of cognitive ability data: a piecewise-constant model and a Weibull model. *Statistics in Medicine* 2008; **27**:5440–5455.
11. Jackson CH, Sharples LD, Thompson SG, Duffy SW. Multistate Markov models for disease progression with classification error. *The Statistician* 2003; **52**(2):193–209.
12. Li YP, Chan W. Analysis of longitudinal multinomial outcome data. *Biometrical Journal* 2006; **48**(2):319–326.
13. Mhoon KB, Chan W, Del Junco DJ, Vernon SW. A continuous-time Markov chain approach analyzing the stages of change construct from a health promotion intervention. *JP Journal of Biostatistics* 2010; **4**(3):213–226.
14. Ma J, Chan W, Tsai C, Xiong M, Tilley B. Analyze Trans-theoretical model of health behavioral changes in a nutrition intervention study – a continuous time Markov chain model with Bayesian approach. *Statistics in Medicine* 2015; **34**:3577–3589.
15. Rosychuk RJ, Thompson ME. Parameter identifiability issues in a latent Markov model for misclassified binary responses. *Journal of the Iranian Statistical Society* 2004; **3**:38–57.
16. Chen B, Yi GY, Cook RJ. Analysis of interval-censored disease progression data via multi-state models under ignorable inspection process. *Statistics in Medicine* 2010; **29**:1175–1189.
17. Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics* 1970; **41**(1):164–171.
18. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA work group under the auspices of department of health and human services task force on Alzheimer's disease. *Neurology* 1984; **34**(7):939–944.
19. Doody RS, Dunn JK, Huang E, Azher S, Kataki M. A method for estimating duration of illness in Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders* 2004; **17**(1-2):1–4.
20. Doody R, Pavlik V, Massman P, Kenan M, Yeh S, Powell S, *et al.* Changing patient characteristics and survival experience in an Alzheimer's center patient cohort. *Dementia and Geriatric Cognitive Disorders* 2005; **20**(2-3):198–208.
21. Folstein MF, Folsetein SE, McHugh PR. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* 1975; **12**(3):189–198.

## Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site.