Published in final edited form as:

Commun Stat Theory Methods. 2019 ; 48(23): 5733–5747. doi:10.1080/03610926.2018.1520884.

Estimation of Multi-state Models with Missing Covariate Values Based on Observed Data Likelihood

Wenjie Lou^{1,3}, Erin L. Abner^{3,4}, Lijie Wan^{1,3}, David W. Fardo^{2,3}, Richard Lipton⁵, Mindy Katz⁵, Richard J. Kryscio^{1,2,3}

¹Department of Statistics, University of Kentucky, Lexington, KY

²Department of Biostatistics, University of Kentucky, Lexington, KY

³Sanders-Brown Center on Aging, University of Kentucky, Lexington, KY

⁴Department of Epidemiology, University of Kentucky, Lexington, KY

⁵Department of Neurology, Albert Einstein College of Medicine, New York City, NY

Abstract

Continuous-time multi-state models are commonly used to study diseases with multiple stages. Potential risk factors associated with the disease are added to the transition intensities of the model as covariates, but missing covariate measurements arise frequently in practice. We propose a likelihood-based method that deals efficiently with a missing covariate in these models. Our simulation study showed that the method performs well for both ‘missing completely at random’ and ‘missing at random’ mechanisms. We also applied our method to a real dataset, the Einstein Aging Study.

Keywords

Longitudinal data; multi-state model; missing covariate; MAR; MCAR

1. Introduction

Survival models are widely used to model processes with two states, often “alive” or “dead.” Continuous-time multi-state models (Commenges, 1999; Hougaard, 1999) are an extension of survival models and are often used to model disease processes that might go through several stages. For example, in the study of dementia, mild cognitive impairment (MCI) is an important intermediate state between normal cognition and dementia; death is an important competing state to MCI and dementia. A multi-state model enables the researcher to study the transitions among these states simultaneously. In clinical applications of multi-state models, one goal is to study the relationship between different covariates and disease evolution. Covariates are often modeled parametrically or semi-parametrically on the

Address for correspondence: Richard J. Kryscio, Room 230, Sanders-Brown Center on Aging, 800 South Limestone Street, Lexington, KY 40536-0230, USA, kryscio@email.uky.edu.

Conflict of Interest

None declared by either author.

transition intensities (Commenges, Joly, Letenneur, & Dartigues, 2004; Kalbfleisch & Lawless, 1985). The Cox Markov multi-state regression model has been widely used in chronic disease studies (Hsieh, Chen, & Chang, 2002; van den Hout & Matthews, 2008).

Several software packages are available for certain types of multi-state models. Hui-Min and colleagues (Hui-Min, Ming-Fang, & Chen, 2004) developed a SAS macro for a non-homogeneous Markov multi-state model, which can be applied to the three-state progressive model. Meira-Machado and colleagues (Meira-Machado, de Una-Alvarez, Cadarso-Suarez, & Andersen, 2009) developed an R library, *tdc.msm*, which can fit 5 different multi-state models, including time-homogeneous Markov models (THMM), non-homogeneous Markov models (NHM), and the Cox Markov model (CMM). Jackson (Jackson, 2011) developed the R package *msm* for time-homogeneous and piece-wise time-homogeneous models.

Even though multi-state models have been widely used in practice, in the current literature there is still no efficient method for handling missing covariate data. The complete case (CC) method is the common approach utilized in most multi-state regression studies and existing software, but there are several limitations to the CC method. CC results in biased estimates of the model parameters when data are not missing completely at random (MCAR). Even when data are MCAR, dropping subjects with missing covariate data can effectively result in loss of oftentimes expensive-to-collect data (Lin & Ying, 1993). Additionally, if there's a large portion of missing data, the CC method may fail due to convergence problems. Like the CC method, the available case (AC) method (Little & Rubin, 2014) is also widely used in practice. With the AC method, any covariate with missing values is removed from the regression model. This method is not helpful when the goal of the analysis is to investigate the effects of these covariates on the response data.

Various imputation methods have also been proposed to deal with missing covariate data. In the unconditional mean substitution method (Little & Rubin, 2014), for example, the missing values are imputed by the unconditional mean based on the observed data. In the conditional mean imputation method (Buck, 1960), a regression model is constructed where the covariate with missing data is the outcome, and the other covariates are the predictors. The regression is fitted on the completely observed data and then the missing covariate values are imputed by linear predictor of the regression model. Both unconditional mean and conditional imputation methods are for missing continuous covariates. To deal with missing data in binary variables, Schemper and Smith (Schemper & Smith, 1990) proposed the probability imputation technique (PIT) for missing prognostic factors in the analysis of clinical trial data. With the PIT method, the missing values are replaced by the sample proportion calculated on the complete cases.

Multiple imputation (Little & Rubin, 2014; Donald B Rubin, 1988; Yuan, 2010) is another popular method for dealing with missing data. Missing data values are estimated by creating multiple complete datasets using methods similar to conditional mean imputation, but here the imputed values are randomly selected from the predictive distribution of the missing values. The predictive distribution is conditioned on all the observed data, including the response data. Each complete dataset is then analyzed and the results combined to estimate the analysis model parameters.

In the present case the observed data contain longitudinal response data with random lengths and unequal spaces, thus it is difficult to identify a predictive model for the covariate with missing data. In practice, an alternative solution is to impute the missing covariate values conditioning only on other observed covariates and ignoring the longitudinal response data. Our simulation and real data application results showed that using multiple imputation methods to impute the missing covariate values based on only the observed covariates could lead to bias.

In this manuscript, we propose a maximum likelihood method to deal with the missing covariate data problem when estimating multi-state models. The method is based on the observed data log-likelihood. For purposes of illustration, we will consider a dataset with one partially missing binary covariate as well as several other covariates with complete measurements. The proposed method works even in situations where the data have mixed discrete-continuous pattern observations (Commenges, 2002), in which the clinical status is assessed at discrete times while the transition time to one of the absorption states (often death) may be observed exactly. If the missing data mechanism is missing at random (MAR), our simulation study showed that the proposed method works well. The method may also be extended to data missing not at random (MNAR) by adding the missing data mechanism to the model

The remainder of this manuscript is organized as follows. First, we review multi-state Markov models with Cox regression intensity functions based on fully observed covariate data in Section 2. In Section 3, we present the proposed method. In Section 4, we present simulation studies to compare the performance of our method to the CC method, the PIT method proposed by Schemper and Smith (Schemper & Smith, 1990), and the MI method based only on the observed covariates. We apply our method to the Einstein Aging Study (EAS) in Section 5. In the concluding section, we discuss the advantages and possible extensions of our method.

2. Multi-state models with complete covariate measurements

Continuous-time multi-state models are based on the theory of multi-state processes. The multi-state process is a stochastic process, $X(t)$, $t \in T$, with a finite state space $S = \{1, 2, \dots, K\}$. It can be characterized fully by its transition intensities (Commenges, 2002). The transition intensity α_{lm} measures the instantaneous hazard of transition from state l to state m . It has the following general form:

$$\alpha_{lm}(t|X(u), u \in [0, t]) = \begin{cases} \lim_{\Delta t \rightarrow 0} \frac{P\{X(t + \Delta t) = m | X(t) = l, X(u), u \in [0, t]\} / \Delta t}{\sum_{k \neq l} \alpha_{lk}(t|X(u), u \in [0, t])} & m \neq l \\ - \sum_{k \neq l} \alpha_{lk}(t|X(u), u \in [0, t]) & m = l \end{cases}$$

2.1. Time-homogenous multi-state process

Although these transition intensities may be dependent on the current time t and the history before time t , $\{X(u), u \in [0, t]\}$, in this study, we assume the process is a time-homogenous Markov process. This implies the transition intensities are independent of the current time t

as well as the history before time t . In this case, the transition probabilities are determined by the matrix of transition intensities $\mathbf{Q}$ with the (l, m) entry α_{lm} .

Let $\mathbf{P}(s, t)$ be the transition probability matrix with the (l, m) entry $p_{lm}(s, t)$, which represents the probability of transition from state l at time s to state m at time t ($s < t$). In the time homogenous case transition probabilities can be calculated by using the forward Kolmogorov equation (Hui-Min et al., 2004), which leads to:

$$\mathbf{P}(s, t) = \mathbf{P}(t - s) = \exp((t - s)\mathbf{Q}).$$

2.2. Cox regression model

In most multi-state model applications, interesting covariates are incorporated into the transition intensities using a Cox form regression model (Andersen et al, 2000; Klein and Shu, 2002):

$$\alpha_{lm}(\mathbf{Z}|\boldsymbol{\beta}) = \alpha_{lm,0} \exp(\boldsymbol{\beta}_{lm}^T \mathbf{Z}) = \exp(\beta_{lm,0} + \boldsymbol{\beta}_{lm}^T \mathbf{Z}) \quad m \neq l.$$

Here $\alpha_{lm,0} = \exp(\beta_{lm,0})$ is the baseline intensity from state l to state m , $\mathbf{Z} = (Z_1, Z_2, \dots, Z_p)$ is a vector of baseline covariates and $\boldsymbol{\beta} = (\beta_{lm,0}, \boldsymbol{\beta}_{lm}; l = 1, \dots, K; m = 1, \dots, K; m \neq l)$, which represents all the parameters associated with this multi-state model.

3. Multi-state models with covariate missingness

In real data applications, it is common that covariates might not be fully observed. In this section, we introduce a likelihood-based method for dealing with partially missing covariate data in the context of continuous-time multi-state models with a continuous-discrete mixed type of observation scheme, as described above (Commenges, 2002). For illustration, we present the case of one partially missing binary covariate along with several other completely observed covariates in the model. The completely observed covariates can be either continuous or discrete.

3.1. The missing data mechanism

First, we define the missing data mechanism (Little & Rubin, 2014; D. B. Rubin, 1976) in this context. Denote $\mathbf{T} = (T_1, T_2, \dots, T_M)$ and $\mathbf{X} = (X_1, X_2, \dots, X_M)$, where M is a random variable indicating the number of observations, T_j is the time of j th observation, and X_j is the corresponding state of the process $X(t)$ occupied at time T_j . We assume that Z_1 is a baseline binary covariate, and its value might not be observed, and $\mathbf{Z}_{obs} = (Z_2, \dots, Z_p)$ is a vector of other baseline covariates whose values are all observed. We define an indicator R such that

$$R = \begin{cases} 1 & \text{if } Z_1 \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

The missing data mechanism can be characterized by the following conditional distribution: $P(R|\mathbf{X}, \mathbf{T}, Z_1, \mathbf{Z}_{obs}, \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ denotes unknown parameters. The data are called MCAR if

the missingness does not depend on the values of X , T , Z_1 and Z_{obs} , that is $P(R|X, T, Z_1, Z_{obs}, \phi) = P(R|\phi)$. The data is called MAR if the missing data mechanism is independent of the underlying value of Z_1 , but might be dependent on the values of T and Z_{obs} , that is $P(R|X, T, Z_1, Z_{obs}, \phi) = P(R|T, Z_{obs}, \phi)$. The data are MNAR if the missing data mechanism is also dependent on the underlying value of Z_1 . In this manuscript, we assume the missing data is MAR.

3.2. Joint modeling of the response and partially missing covariate

Here we assume the observation process is ignorable (Commenges, 2002), thus the time points of the observation process T can be viewed as fixed. The proposed method is based on the conditional joint distribution of $(R, X, Z_1|T, Z_{obs})$. Under the assumption of MAR, the distribution can be written as

$$P(R, X, Z_1|T, Z_{obs}) = P(R|X, T, Z_{obs}, \phi)P(X|T, Z_1, Z_{obs}, \beta)P(Z_1|Z_{obs}, \gamma)$$

(1).

Here, β is the vector of parameters of interest and is associated with the multi-state model. In this manuscript we assume that

$$P(R|X, T, Z_{obs}, \phi) = P(R|T, Z_{obs}, \phi)$$

(2)

which does not depend on X , implying that ϕ is a vector of nuisance parameters. The complexities that can arise when this conditional probability depends on $\mathbf{X}$ are discussed by Wang et al. (2002).

Since Z_1 is binary, we can use logistic regression to model the distribution of Z_1 :

$$\log\left(\frac{p_1}{1-p_1}\right) = \gamma_0 + \phi^T Z_{obs}$$

Here, $p_1 = P(Z_1 = 1|Z_{obs}, \gamma)$ and $\gamma = (\gamma_0, \phi)$. In situations where Z_1 and Z_{obs} are independent, or there's just one covariate Z_1 and no Z_{obs} variables, we could have

$$\log\left(\frac{p_1}{1-p_1}\right) = \gamma_0.$$

With the same method, we can generalize the model to the case when Z_1 has more than two levels or we have multiple categorical covariates with missing data. In that case, we replace

the above binary logistic model with a multinomial logistic model. When the covariate has missing values but is a continuous variable, then the logistic model is replaced by a probability density function (see the Discussion section).

3.3. The likelihood function

Assume that states 1 to $K-2$ are transient states, and the last two states are absorbing (e.g. dementia and death before a dementia as discussed in Figure 1 below). Since we cannot make observations in continuous time but rather only at a finite number of distinct times, the exact transition times to state 1, 2, ..., $K-2$ are unknown and are all interval-censored. The exact time of death can be retrieved, but the state just before death is unknown. Due to the presence of the absorbing states, the number of observations varies by subject. Let m_i denote the number of observations for subject i .

Denote $\mathbf{t}_i = (t_{i,1}, t_{i,2}, \dots, t_{i,m_i})$ and $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,m_i})$, where $t_{i,j}$ is the time point of the j th observation for subject i , and $x_{i,j}$ is the corresponding occupied state at that time point. Define index of death $\delta_i = I(x_{i,m_i} = K)$. Thus we have $\delta_i = 1$ if subject i died at time t_{i,m_i} , and $\delta_i = 0$ otherwise. Denote the covariates $\mathbf{z}_i = (z_{i1}, \mathbf{z}_{i,obs})$, where $\mathbf{z}_{i,obs} = (z_{i2}, \dots, z_{ip})$. The likelihood for the i th subject, L_i , can be calculated from Equations (1) and (2) as follows:

$$L_i = P(r_i | \mathbf{t}_i, \mathbf{z}_{i,obs}, \boldsymbol{\phi}) \times \left(P(\mathbf{x}_i | \mathbf{t}_i, \mathbf{z}_i, \boldsymbol{\beta}) P(\mathbf{z}_{i,1} | \mathbf{z}_{i,obs}, \boldsymbol{\gamma}) \right)^{r_i} \left(\sum_{z_{i,1}=0}^1 P(\mathbf{x}_i | \mathbf{t}_i, \mathbf{z}_i, \boldsymbol{\beta}) P(\mathbf{z}_{i,1} | \mathbf{z}_{i,obs}, \boldsymbol{\gamma}) \right)^{1-r_i}.$$

Here $r_i = 1$ if z_{i1} is observed and 0 otherwise, and the conditional distribution $P(\mathbf{x}_i | \mathbf{t}_i, \mathbf{z}_i, \boldsymbol{\beta})$, under the assumption that the process is Markov and given the baseline state, can be calculated as:

$$P(\mathbf{x}_i | \mathbf{t}_i, \mathbf{z}_i, \boldsymbol{\beta}) = \prod_{j=2}^{m_i} P(x_{i,j} | x_{i,j-1}, t_{i,j-1}, t_{i,j}, \mathbf{z}_i, \boldsymbol{\beta}).$$

Here,

$$P(x_{i,j} | x_{i,j-1}, t_{i,j-1}, t_{i,j}, \mathbf{z}_i, \boldsymbol{\beta}) = \begin{cases} P_{x_{i,j-1}, x_{i,j}}(t_{i,j-1}, t_{i,j} | \mathbf{z}_i, \boldsymbol{\beta}) & \text{if } j \neq m_i \\ \left[P_{x_{i,j-1}, x_{i,j}}(t_{i,j-1}, t_{i,j} | \mathbf{z}_i, \boldsymbol{\beta}) \right]^{1-\delta_i} \left[\sum_{k \neq K} P_{x_{i,j-1}, k}(t_{i,j-1}, t_{i,j} | \mathbf{z}_i, \boldsymbol{\beta}) \alpha_{kK}(\mathbf{z}_i | \boldsymbol{\beta}) \right]^{\delta_i} & \text{if } j = m_i \end{cases}$$

The log-likelihood function of all the subjects can be written as

$$\ell = l(\phi) + l(\beta, \gamma).$$

Here

$$l(\phi) = \sum_{i=1}^n \log(P(r_i | t_i, z_{i, obs}, \phi))$$

and

$$l(\beta, \gamma) = \sum_{i=1}^n \left(r_i \log(P(x_i | z_i, \beta) P(z_{i,1} | z_{i, obs}, \gamma)) + (1 - r_i) \log \left(\sum_{z_{i,1}=0}^1 P(x_i | z_i, \beta) P(z_{i,1} | z_{i, obs}, \gamma) \right) \right).$$

If ϕ is not a function of (β, γ) , then we have $\ell \propto l(\beta, \gamma)$. The Newton-Raphson method can be used to maximize the log-likelihood function $l(\beta, \gamma)$.

4. Simulation study

In this section, we evaluate the performance of the proposed method (referred to as OL throughout) through simulations. Both MCAR and MAR data will be considered here. We analyzed the full dataset with no missing values, that is, prior to setting any values to missing (referred to as FULL). This estimation acts like a benchmark, so we can evaluate the performance of our method in general. To assess advantages of the method, we also compared it to the CC method, the PIT method proposed by Schemper and Smith (Schemper & Smith, 1990), and the MI method. In applying the MI method, we imputed the missing covariate values by ignoring the longitudinal response data and only conditioning on other fully observed covariates. The criteria used in the comparisons are percent bias (% bias), estimated standard error (SE) and 95% confidence interval coverage probability (95% CP). The results are based on 500 simulated datasets. All simulations were done in SAS 9.3® (SAS Institute, 2011). PROC IML was used to calculate and maximize the likelihood functions. PROC MI and PROC MIANALYZE were used to impute datasets and to combine results for the MI method.

We considered two covariates (Z_1, Z_2) in this simulation study. The first covariate Z_1 is a baseline binary covariate, and its value is missing for some subjects. The second covariate Z_2 is a continuous covariate, and its value is completely observed. We generated Z_2 from a normal distribution with $(\mu = 1, \sigma = 0.25)$ and truncated at 0 and 2. We generated Z_1 from a logistic regression model based on the value of Z_2 , which has the following form

$$P(Z_1 = 1 | Z_2) = \frac{\exp(\gamma_0 + \gamma_1 Z_2)}{1 + \exp(\gamma_0 + \gamma_1 Z_2)}.$$

We set $(\gamma_0, \gamma_1) = (-0.2, 0.2)$ in the true model.

The longitudinal response data were generated from a three-state model with the following transition intensity matrix:

$$Q = \begin{pmatrix} -(0.2 + \exp(-2.2 + \beta_1 Z_1)) & \exp(-2.2 + \beta_1 Z_1) & 0.2 \\ 0 & -\exp(-2.0 + \beta_2 Z_2) & \exp(-2.0 + \beta_2 Z_2) \\ 0 & 0 & 0 \end{pmatrix}$$

The true values of β_1 and β_2 are 0.8 and 0.4 respectively. Subjects are in state 1 at time 0. States 1 and 2 are transient states, while state 3 is an absorbing state. The states occupied were observed at 1-year intervals with a common censoring time of 10 years if the subject remained in state 1 or state 2. The number of subjects (N) was 1000.

We considered both MCAR and MAR mechanisms in this study. For MCAR data, defined by $P(R|T, Z_{obs}, \phi) = P(R|\phi)$, we considered two cases. In the first case, we randomly set about 30% of the observed values of Z_1 to missing. We denote this case as MCAR 1. In the second case, we randomly set Z_1 to missing for about 70% subjects. We denote this case as MCAR 2. For MAR data, defined by $P(R|T, Z_{obs}, \phi) = P(R|T, \phi)$, we also considered two cases. In the first case, we set Z_1 missing if the first transition to states 2 or 3 happened within the first year. In this situation, we have about 30% missingness in the covariate. We denote this case as MAR 1. In the second case, we set the observed value of Z_1 missing if the first transition happened within the first 3 years. In this case, we have about 70% missingness in the covariate. We denote this case as MAR 2. These missing data are MAR, since whether the covariate is missing only depends on the observed data, which is the transition history, but not on the unobserved data, which is the underlying value of Z_1 .

The simulation results are listed in Table 1. When the missing data were MCAR, the CC, PIT and OL methods all worked well. The percent bias was small, and the real coverage rates for the nominal 95% confidence intervals were all near 95%. Both results were close to those provided by the full data estimates in which we had completely observed covariate values for all subjects. The PIT method had relatively larger bias than both the CC method and the proposed OL method for the coefficient estimate of the partially missing covariate Z_1 , but the bias is in an acceptable range. The CC method was less efficient than both the PIT method and the proposed OL method in terms of standard errors, especially for the MCAR 2 case, which had a higher missing proportion. The MI method performed the worst among the four methods considered. The results by MI method had large bias for the coefficient estimate of the partially missing covariate Z_1 even though the data were MCAR.

When the data were MAR, the proposed OL method still worked well, while all other methods considered here had relatively large biases. The CC method had as large as 120% bias for the coefficient estimate of Z_1 . The PIT and MI method performed relatively better than the CC method, but both methods had at least 10% bias for the coefficient estimate of Z_1 whether Z_1 and Z_2 were independent or not. However, our method (OL) still worked well for MAR data; the estimates provided by our method were unbiased. The simulation results also showed that our method did not lose much efficiency because of the missing data. The

increases of the standard errors were relatively small compared to those for the full data estimates (FULL), where we had complete covariate data for all subjects.

5. Application

The Statistical Modeling of Aging and Risk of Transition (SMART) project is a multi-center dementia study. This project aggregates data from 11 mature and well-known longitudinal cohorts of older adults with high autopsy rates (Abner et al., 2015). In this application, we used data from the Einstein Aging Study (EAS), which contributed a set of standardized data to the SMART database. EAS, located at the Albert Einstein College of Medicine, Yeshiva University, is a population-based study of cognitive aging and dementia in a non-institutionalized, urban, and multi-ethnic community. Participants undergo annual in-person evaluations that include cognitive, neurological, functional, and physical assessments (Katz et al., 2012).

Three cognitive states were of interest: Cognitively Intact (state 1), MCI (state 2), and Dementia (state 3). Death (state 4) was also included in our model as a competing risk for MCI and Dementia. Cognitively Intact, MCI, and Dementia were transient states, while Death was an absorbing state. Transitions from Cognitively Intact to MCI, Dementia, and Death were possible. Transitions from MCI to Cognitively Intact, Dementia, and Death were possible. No transitions to better cognitive states were possible from Dementia, but transition from Dementia to Death was possible. See Figure 1 for the transition model structure.

The original dataset contained longitudinally observed data on 2,097 patients. In this application, we excluded participants if they had only one observation ($n=500$) or already entered into MCI or Dementia ($n=395$) at EAS baseline. The 1,202 participants included in this study were in the Cognitively Intact state at EAS baseline, and there were 4,302 total observed transitions. This resulted in 4.6 ± 2.7 transitions per subject with a mean follow-up 5.7 ± 3.6 years. See Table 2 for the observed transition frequencies for both the full data and the dataset with only the complete cases.

One covariate of interest for our application is APOE, a genetic marker of Alzheimer's disease risk; APOE genotype was not available for all EAS participants. There were 509 (42.4%) participants with missing APOE (APOE4: any 4 allele=1, no 4 alleles=0). Other risk factors of interest in this model are baseline age (Age), sex (Female: 1=female, 0=male) and education level. Education level (LowEdu) was dichotomized into two groups, low education (≤ 12 years) and high education (> 12 years). Baseline age, education level, and sex were all fully observed. Summaries of these risk factors are listed in Table 3. APOE4 carrier status is the single most important genetic predictor of cognitive impairment and dementia, and the most important socio-demographic variables are age, sex, and years of education (Plassman et al., 2011).

We used a time-homogenous Markov model to study the effects of these risk factors on each transition. This four-state model has the following transition intensity matrix:

$$Q(\mathbf{Z}) = \begin{bmatrix} \alpha_{11}(\mathbf{Z}) & \alpha_{12}(\mathbf{Z}) & \alpha_{13}(\mathbf{Z}) & \alpha_{14}(\mathbf{Z}) \\ \alpha_{21}(\mathbf{Z}) & \alpha_{22}(\mathbf{Z}) & \alpha_{23}(\mathbf{Z}) & \alpha_{24}(\mathbf{Z}) \\ 0 & 0 & \alpha_{33}(\mathbf{Z}) & \alpha_{34}(\mathbf{Z}) \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Here $\mathbf{Z}$ is the vector of above risk factors, namely Bage, Female, LowEdu, and APOE4. Since all participants were over 60 years old at baseline, we centered the original baseline age at 60. Since the effects of risk factors on the backward transition from MCI to Cognitively Intact were not of interest in this study, we didn't add covariates on the transition intensity function for this path. Thus, the non-zero intensity functions in the transition intensity matrix $Q(\mathbf{Z})$ have the following Cox-type regression form:

$$\alpha_{lm}(\mathbf{Z}) = \begin{cases} \exp(\beta_{lm,0}) & \text{if } l = 2, m = 1 \\ \exp(\beta_{lm,0} + \beta_{lm}^T \mathbf{Z}) & \text{if } l < m \\ - \sum_{h \neq l} \alpha_{lh}(\mathbf{Z}) & \text{if } l = m \end{cases}$$

We considered five methods to deal with the missing APOE data: the AC method, the CC method, the PIT method, the proposed OL method, and the MI method, for which we ignored the longitudinal response data. For each method, a backwards algorithm was used for model selection. A basic model was fitted to the data with all risk factors modeled on all possible transitions. At each step the coefficient with the largest p value was eliminated from the model until all coefficients remaining in the model were significant at the 0.05 level.

First, we fitted the model without APOE4. With APOE4 excluded from the model, the data were completely observed (AC method). Table 4 lists the hazard ratios and the corresponding 95% confidence intervals for all the significant risk factors ($P < 0.05$). We used these results to evaluate how the missing APOE4 data affected the results for the CC, PIT, our proposed OL method and the conventional MI method by imputing the missing APOE4 values based on only the other observed covariates.

For the proposed OL method, the distribution of APOE4 is constructed by the logistic regression model based on the fully observed covariates Bage, Female and LowEdu, which has the following form:

$$\log\left(\frac{P(\text{APOE4} = 1)}{1 - P(\text{APOE4} = 1)}\right) = \varphi_0 + \varphi_1 \text{Bage} + \varphi_2 \text{Female} + \varphi_3 \text{LowEdu}.$$

The results generated by the OL method are listed in Table 5.

Table 6 lists the results by using the PIT method. In the PIT method, missing APOE4 values were imputed by its percentage based the observed data, which was 21.6% in our data (see Table 3).

Since the likelihood function contained as many as 30 parameters in the initial model, and 42.4% of the subjects had missing data, the model failed to converge using the CC method. We refitted the CC method on the reduced model based on the results of the PIT method and the proposed OL method. The results are listed in Table 7.

When applying the MI method, we imputed the missing APOE4 values via logistic regression using the fully observed covariates Bage, Female and LowEdu as predictors. The pooled MI results showed that APOE4 was not significant on any of the transition paths. Thus, the final model reduced to same model as the AC method. In order to compare the MI method to the proposed OL method, we refitted the MI method on the reduced model based on the results of the PIT method and the OL method. The results are listed in Table 8.

The PIT method and the proposed OL method provided very similar results. The parameter estimates for baseline age, female sex, and low education on each path by both methods were also very close to those produced by the AC method. In general, a 1-year increase in baseline age increased the transition intensities from Cognitively Intact to MCI and MCI to Dementia. Cognitively Intact women had a lower mortality rate than Cognitively Intact men. Cognitively Intact subjects with a low education level had a higher risk for the transition to MCI. Additionally, both the PIT method and the OL method showed that APOE4 increased the transition intensities from Cognitively Intact to MCI. These findings are much in line with other similar studies (Kryscio et al., 2013; Song, Kuo, Derby, Lipton, & Hall, 2011). The CC method failed to indicate that female sex had significant effects on the transition intensity from Cognitively Intact to Death, while the effects of baseline age, low education level, and APOE4 are in agreement with the results by PIT method and OL method. The MI method, by ignoring the longitudinal outcome data, failed to show APOE4 was significant.

We used the length of the 95% confidence intervals to compare the efficiency among the methods which were considered here. The results showed that the CC method was the least efficient method among the three. It had the widest 95% confidence intervals. The proposed OL method showed a small advantage over the PIT method, such that the width of the 95% confidence interval of the hazard ratio for APOE4 on the path from Cognitively Intact to MCI was narrower than that yielded by the PIT method (0.66 vs. 0.70). The MI method considered here had similar efficiency to the proposed OL method. However, the MI method failed to identify APOE4 as a significant predictor.

6. Discussion

Multi-state models have become a popular tool to study transitions in studies of chronic disease. But, most model estimation methods and applications require fully observed data. We proposed a likelihood-based method that would handle missing covariate data in continuous-time Markov multi-state models. We demonstrated that the proposed method works for datasets with one missing categorical covariates as with several other completely observed covariates. The completely observed covariates are allowed to be either continuous or discrete. Other applications of the proposed method are described below.

Compared to the CC method, which is commonly used in multi-state models with missing data, our method had three major advantages. First, when the data were MCAR our method was more efficient, especially when the data contained a large portion of cases with missing data. Second, our method worked for both MCAR data and MAR data, while the CC method may provide biased estimates if the data are not MCAR. Third, when the dataset contained a large portion of cases with missing covariate values, the CC method might fail due to a convergence problem. In this case, the proposed method would be an alternative for the analysis.

From our simulation studies, the PIT method proposed by Schemper and Smith may also produce biased results when the data are not MCAR. The covariate distributions of the missing covariates for the missing cases and observed cases would be different when the missingness is dependent on other observed data. Thus, the PIT method is not recommended when the data is MAR not MCAR in estimation of multi-state models with missing categorical covariate.

The MI method has become a widely used tool to deal with missing data. It is an efficient method and is also easy to carry out in many statistical modeling applications. However, in the area of multi-state models, the proper MI method is often not easy to carry out. Our study showed that the conventional method of imputing the missing covariate values by a model based only on the other fully observed covariates is biased even when the data are MCAR. Similar comments of MI method by ignoring the response data can also be found in Little (Little, 1992) and Lin and Ying (Lin & Ying, 1993).

In our application to the EAS dataset, CC had the widest CIs and was not able to identify female sex as protective for transitions from Cognitively Intact to Death. The OL and PIT methods gave the same results in the application, which was within our expectation since the missing APOE in the EAS dataset is likely to have arisen more from an MCAR mechanism (i.e., administratively based reasons) than MAR (i.e., participant based reasons). The MI method based on the fully observed covariates showed bias in the estimation of the effect of APOE4 in the model.

The proposed method could be extended to other scenarios. If the covariate with missing data is continuous, the observed data log-likelihood involves integration. In this circumstance calculation of the log-likelihood would be difficult, since there are no closed form solutions for the integration. However, numerical integration methods, such as Gaussian quadrature, Monte Carlo simulations, quasi-Monte Carlo, etc., can be used to approximate observed data log-likelihood.

The proposed method could also be extended to data with multiple categorical covariates with missing data. However, if the data contain too many covariates with missing values, this method might encounter programming difficulties or convergence problems when fitting the model. Since an additional model is needed for all the covariates with missing values, modeling multiple covariates with missing data might introduce many nuisance parameters into the likelihood function. With too many nuisance parameters in the likelihood function, there might be difficulties computing estimates using the usual maximization techniques,

like the Newton-Raphson method. One possible solution to the problem discussed above is to use the EM algorithm as discussed by Ibrahim et al. (Ibrahim, Chen, Lipsitz, & Herring, 2005). By applying the EM algorithm, the nuisance parameters associated with the covariates model can be estimated separately from the model parameters associated with the multi-state model. The authors of this paper are pursuing this method.

Acknowledgments

Funding

This research was partially funded with support from the following grants to the University of Kentucky's Center on Aging: R01 AG038651, P30 AG028383, K25 AG043546, and R01 AG019241 from the National Institute on Aging, as well as a grant to the University of Kentucky's Center for Clinical and Translational Science, UL1TR000117, from the National Center for Advancing Translational Sciences. The Einstein Aging Study is supported by P01 AG003949.

References

- Abner EL, Schmitt F, Nelson P, Lou W, Wan L, Gauriglia R,... Kryscio RJ. (2015). The Statistical Modeling of Aging and Risk of Transition Project: Data Collection and Harmonization Across 11 Longitudinal Cohort Studies of Aging, Cognition, and Dementia. *Obs Stud*, 1(2015), 56. [PubMed: 25984574]
- Andersen PK, Esbjerg S, and Sorensen TIA, Multi-state models for bleeding episodes and mortality in liver cirrhosis. *Statistics in Medicine*, 2000 19(4): p. 587–599. [PubMed: 10694738]
- Buck SF (1960). A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic-Computer. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 22(2), 302–306.
- Commenges D (1999). Multi-state models in epidemiology. *Lifetime Data Anal*, 5(4), 315–327. [PubMed: 10650740]
- Commenges D (2002). Inference for multi-state models from interval-censored data. *Stat Methods Med Res*, 11(2), 167–182. [PubMed: 12040695]
- Commenges D, Joly P, Letenneur L, & Dartigues JF (2004). Incidence and mortality of Alzheimer's disease or dementia using an illness-death model. *Stat Med*, 23(2), 199–210. doi:10.1002/sim.1709 [PubMed: 14716722]
- Hougaard P (1999). Multi-state models: a review. *Lifetime Data Anal*, 5(3), 239–264. [PubMed: 10518372]
- Hsieh HJ, Chen TH, & Chang SH (2002). Assessing chronic disease progression using non-homogeneous exponential regression Markov models: an illustration using a selective breast cancer screening in Taiwan. *Stat Med*, 21(22), 3369–3382. doi:10.1002/sim.1277 [PubMed: 12407678]
- Hui-Min W, Ming-Fang Y, & Chen TH-H (2004). SAS macro program for non-homogeneous Markov process in modeling multi-state disease progression. *Comput Methods Programs Biomed*, 75(2), 95–105. [PubMed: 15212852]
- Ibrahim JG, Chen MH, Lipsitz SR, & Herring AH (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469), 332–346. doi:10.1198/016214504000001844
- Institute S (2011). *SAS/IML 9.3 User's Guide*: SAS Institute.
- Jackson CH (2011). Multi-State Models for Panel Data: The msm Package for R. *Journal of Statistical Software*, 38(8), 1–28.
- Kalbfleisch JD, & Lawless JF (1985). The Analysis of Panel Data under a Markov Assumption. *Journal of the American Statistical Association*, 80(392), 863–871. doi:10.2307/2288545
- Katz MJ, Lipton RB, Hall CB, Zimmerman ME, Sanders AE, Verghese J,... Derby CA (2012). Age-specific and Sex-specific Prevalence and Incidence of Mild Cognitive Impairment, Dementia, and Alzheimer Dementia in Blacks and Whites A Report From the Einstein Aging Study. *Alzheimer*

Disease & Associated Disorders, 26(4), 335–343. doi: 10.1097/WAD.0b013e31823dbcf6 [PubMed: 22156756]

- Klein JP and Shu YY. (2002) Multi-state models for bone marrow transplantation studies. *Statistical Methods in Medical Research*, 11(2):. 117–139. [PubMed: 12040693]
- Kryscio RJ, Abner EL, Lin YS, Cooper GE, Fardo DW, Jicha GA,... Schmitt FA (2013). Adjusting for Mortality when Identifying Risk Factors for Transitions to Mild Cognitive Impairment and Dementia. *Journal of Alzheimers Disease*, 35(4), 823–832. doi: 10.3233/Jad-122146
- Lin DY, & Ying Z (1993). Cox Regression with Incomplete Covariate Measurements. *Journal of the American Statistical Association*, 88(424), 1341–1349. doi:10.2307/2291275
- Little RJ (1992). Regression with missing X's: a review. *Journal of the American Statistical Association*, 87(420), 1227–1237.
- Little RJ, & Rubin DB (2014). *Statistical analysis with missing data*: John Wiley & Sons.
- Meira-Machado L, de Una-Alvarez J, Cadarso-Suarez C, & Andersen PK (2009). Multi-state models for the analysis of time-to-event data. *Stat Methods Med Res*, 18(2), 195–222. doi: 10.1177/0962280208092301 [PubMed: 18562394]
- Plassman BL, Langa KM, McCammon RJ, Fisher GG, ..., Wallace RB. (2011) Incidence of dementia and cognitive impairment not demenita in the United States. *Ann Neurol* 70: 418–426. [PubMed: 21425187]
- Rubin DB (1976). Inference and Missing Data. *Biometrika*, 63(3), 581–590. doi: DOI 10.1093/biomet/63.3.581
- Rubin DB (1988). An overview of multiple imputation Paper presented at the Proceedings of the survey research methods section of the American statistical association.
- Schemper M, & Smith TL (1990). Efficient Evaluation of Treatment Effects in the Presence of Missing Covariate Values. *Stat Med*, 9(7), 777–784. doi:10.1002/sim.4780090707 [PubMed: 2218180]
- Song C, Kuo L, Derby CA, Lipton RB, & Hall CB (2011). Multi-stage transitional models with random effects and their application to the Einstein aging study. *Biometrical Journal*, 53(6), 938–955. [PubMed: 22020750]
- van den Hout A, & Matthews FE (2008). Multi-state analysis of cognitive ability data: a piecewise-constant model and a Weibull model. *Stat Med*, 27(26), 5440–5455. doi: 10.1002/sim.3360. [PubMed: 18618421]
- Wang CY, Chen JC, Lee S-M, Ou ST (2002) Joint conditoinal likelihood estimator in logistic regression with missing covariate data. *Statistica Sinica*, 12: 555–574.
- Yuan YC (2010). *Multiple imputation for missing data: Concepts and new development (Version 9.0)*. SAS Institute Inc, Rockville, MD, 49.

Figure 1.
Multi-state model structure of the 4 states.

Table 1:

Simulation Results for datasets in where Z1 and Z2 are dependent

MissMech	Method	$\beta_1 (0.8)$			$\beta_2 (0.4)$		
		% bias	SE	95% CP	% bias	SE	95% CP
None	FULL	0.15%	0.064	94.6%	1.61%	0.056	94.4%
	CC	0.27%	0.077	94.6%	2.05%	0.067	95.0%
MCAR1	PIT	0.74%	0.071	93.8%	1.51%	0.056	94.0%
	OL	0.22%	0.071	94.2%	1.59%	0.056	94.0%
	MI	-14.37%	0.073	69.8%	-0.01%	0.056	94.8%
	CC	0.59%	0.118	95.4%	2.54%	0.102	93.6%
	PIT	1.96%	0.084	92.2%	1.42%	0.056	94.2%
MCAR2	OL	0.31%	0.085	94.4%	1.57%	0.056	94.4%
	MI	-33.78%	0.083	6.20%	-2.13%	0.056	93.8%
	CC	-51.56%	0.081	0.00%	-4.89%	0.071	93.4%
	PIT	-12.15%	0.069	73.8%	-0.80%	0.056	95.2%
	OL	0.23%	0.068	94.2%	1.65%	0.056	94.0%
MAR1	MI	-13.33%	0.074	75.4%	-0.43%	0.056	95.0%
	CC	-118.32%	0.128	0.00%	-11.41%	0.119	93.8%
	PIT	-18.36%	0.082	57.8%	-2.74%	0.056	94.0%
	OL	0.13%	0.072	94.8%	1.60%	0.056	94.4%
	MI	-30.65%	0.091	17.6%	-3.14%	0.056	93.6%

Note: MissMech=Missing Mechanism. FULL = analysis of fully observed data. CC = complete case method. PIT = probability imputation technique. OL = observed data likelihood. MI = multiple imputation. MCAR = missing completely at random, MAR = missing at random. MCAR 1 & MAR 1 denote simulated data with 30% missing values for Z1, and MCAR2 & MAR2 denote simulated data with 70% missing values for Z1.

Table 2:

Observed frequency of transitions in the Einstein Aging Study dataset (N=1,202 participants) (Left: Full data; Right: Complete Case).

From	To				To			
	INT	MCI	Dementia	Death	INT	MCI	Dementia	Death
INT	3027	328	37	454	2260	236	24	185
MCI	151	126	42	47	121	97	32	21
Dementia			48	42			30	27

Note: INT = Cognitively Intact; MCI = Mild Cognitive Impairment.

Table 3:

Summary statistics of baseline risk factors in Einstein Aging Study participants (N=1,202).

Baseline Risk Factor	N Missing (%)	N Percent (%) or mean $\pm$ st. dev.
Baseline age, years	0 (0)	78.88 $\pm$ 5.34
Female sex	0 (0)	724 (60.2)
Low Education ($\leq$ 12 years)	0 (0)	593 (49.3)
APOE4 (ϵ 4 allele)	509 (42.4)	150 (21.6)

Table 4:

Hazard ratios (HR) for risk factors using Available Case method.

Risk Factor	Path	HR	95% Confidence Interval		
			L	U	Length
Baseline Age	INT to MCI	1.07	1.05	1.08	0.03
	MCI to Dementia	1.04	1.00	1.07	0.07
Female	INT to Death	0.43	0.22	0.86	0.64
Low Education	INT to MCI	1.54	1.30	1.82	0.52

Note: INT = Cognitively Intact. MCI = Mild Cognitive Impairment.

Table 5:

Hazard ratios (HR) for risk factors using Observed Data Likelihood method.

Risk Factor	Path	HR	95% Confidence Interval		
			L	U	Length
Baseline Age, 1-year	INT to MCI	1.07	1.05	1.08	0.03
	MCI to Dementia	1.04	1.01	1.07	0.06
Female	INT to Death	0.45	0.24	0.85	0.61
Low Education	INT to MCI	1.54	1.30	1.83	0.53
APOE4	INT to MCI	1.44	1.15	1.81	0.66

Note: INT = Cognitively Intact. MCI = Mild Cognitive Impairment.

Table 6:

Hazard ratios (HR) for risk factors using Probability Imputation Technique

Risk Factor	Path	HR	95% Confidence Interval		
			L	U	Length
Baseline Age, 1-year	INT to MCI	1.07	1.05	1.09	0.03
	MCI to Dementia	1.04	1.01	1.07	0.06
Female	INT to Death	0.44	0.23	0.84	0.61
Low Education	INT to MCI	1.54	1.30	1.83	0.53
APOE4	INT to MCI	1.47	1.16	1.86	0.70

Note: INT = Cognitively Intact. MCI = Mild Cognitive Impairment.

Table 7:

Hazard ratios (HR) for risk factors using Complete Case method on the reduced model

Risk Factor	Path	HR	95% Confidence Interval		
			L	U	Length
Baseline Age, 1-year	INT to MCI	1.07	1.05	1.09	0.04
	MCI to Dementia	1.05	1.01	1.10	0.09
Female	INT to Death	0.25	0.05	1.25	1.20
Low Education	INT to MCI	1.59	1.29	1.97	0.68
APOE4	INT to MCI	1.52	1.19	1.95	0.76

Note: INT = Cognitively Intact. MCI = Mild Cognitive Impairment.

Table 8:

Hazard ratios (HR) for risk factors using Multiple Imputation method on the reduced model

Risk Factor	Path	HR	95% Confidence Interval		
			L	U	Length
Baseline Age	INT to MCI	1.07	1.05	1.09	0.03
	MCI to Dementia	1.04	1.01	1.07	0.07
Female	INT to Death	0.43	0.22	0.85	0.62
Low Education	INT to MCI	1.54	1.30	1.82	0.53
APOE4	INT to MCI	1.25	0.98	1.56	0.58

Note: INT = Cognitively Intact. MCI = Mild Cognitive Impairment.