Published in final edited form as:

Biom J. 2011 May ; 53(3): 444–463. doi:10.1002/bimj.201000122.

Non-homogeneous Markov process models with informative observations with an application to Alzheimer's disease

Baojiang Chen^{and*}1 and Xiao-Hua Zhou²

¹Department of Biostatistics, University of Nebraska Medical Center, Omaha, NE 68198, USA

²Department of Biostatistics, University of Washington, Seattle, WA 98105, USA

Abstract

Identifying risk factors for transition rates among normal cognition, mildly cognitive impairment, dementia and death in an Alzheimer's disease study is very important. It is known that transition rates among these states are strongly time dependent. While Markov process models are often used to describe these disease progressions, the literature mainly focuses on time homogeneous processes, and limited tools are available for dealing with non-homogeneity. Further, patients may choose when they want to visit the clinics, which creates informative observations. In this paper, we develop methods to deal with non-homogeneous Markov processes through time scale transformation when observation times are pre-planned with some observations missing. Maximum likelihood estimation via the EM algorithm is derived for parameter estimation. Simulation studies demonstrate that the proposed method works well under a variety of situations. An application to the Alzheimer's disease study identifies that there is a significant increase in transition rates as a function of time. Furthermore, our models reveal that the non-ignorable missing mechanism is perhaps reasonable.

Keywords

Markov; Missing data; Non-homogeneous; Transformation

1 Introduction

In studies of Alzheimer's disease (AD), mildly cognitive impairment (MCI) is recognized as an important transitional state between normal cognition and dementia. One of the objectives of the study is to identify risk factors for transitions among these states. In clinical studies, cognitive status is often assessed only at periodic follow-up visits, with exact times of transition and trajectory of states visited between observations unknown, resulting in interval-censored data. Multi-state models provide a convenient way to characterize the movement of individuals among distinct states. With continuous time multi-state models, transition intensities are often modeled using Markov models (e.g. Bartholomew, 1983; Singer and Spilerman, 1976 a, b; Wasserman, 1980). Of course, with estimated transition densities, we can easily compute transition probabilities. Various methods based on Markov models have been proposed in literature, including discrete time (Albert and Waclawi, 1998) and continuous time models (Andersen et al., 1993). For continuous time models,

Conflict of Interest

The authors have declared no conflict of interest.

Supporting Information for this article is available from the author or on the WWW under <http://dx.doi.org/10.1002/bimj.201000122>

when observation times are not evenly spaced, Kalbfleisch and Lawless (1985, 1989) proposed an effective method for time homogeneous processes. Cook et al. (2004) described a conditional Markov model with multivariate random effects to handle clustered, conditionally Markov, multi-state processes. Motivated by studies of smoking behavior, Cook et al. (2002) developed other extensions of Markov models, which accommodate heterogeneity in the patterns of movement between states by allowing subject-specific absorbing states.

Most applications assume a homogeneous process; that is, the transition probabilities only depend on the elapsed time between observations. This assumption is not satisfied when transition probabilities depend on time from the process origin. The general method to deal with a homogeneous model is straightforward, and limited work has been devoted to addressing non-homogeneity in Markov process models. Kalbfleisch and Lawless (1985) proposed a method for modeling non-homogeneous multi-state processes under panel observations, in which the non-homogeneous intensity matrix is a product of a baseline homogeneous intensity matrix and a function of time. Gentleman et al. (1994) considered piecewise constant transition intensities to accommodate non-homogeneity in multi-state models. A number of authors have used piecewise constant transition intensities to model temporal homogeneity with applications (e.g. Ocana-Riola, 2005; Perez-Ocon et al., 2001; Saint-Pierre et al., 2003). Hubbard et al. (2008) considered a time-transformation method to address the non-homogeneity. This method allows for variation in transition intensities by assuming that the time-varying transition intensity matrix is a product of a baseline homogeneous transition intensity matrix and a scalar function of time. For the other methods, interested reader can refer to Cappe et al. (2005), and Breto et al. (2009).

In cohort studies, clinical assessments may be scheduled before the study, but patients may choose when they want to visit clinics for clinical examinations according to their degree of disease activity. This creates a problem somewhat akin to incomplete data arising in longitudinal studies. In this case, data may be missing at random (MAR) (Laird, 1988; Little and Rubin, 2002) if missing status depends on observed (typically past) responses, or missing not at random (MNAR), where the missing status may depend on the latent disease status. For example, in AD studies, patients may choose to visit the clinics according to their past observed disease status or the present disease status, which creates the MAR or MNAR mechanism.

Little work in literature has addressed informative observations under the framework of non-homogeneous Markov processes. Under an MAR or MNAR mechanism, the naive analysis method such as the complete case analysis can give biased inferences. Grüger et al. (1991) discussed the informative sampling in multi-state models. Chen et al. (2010) proposed a piecewise constant transition model to address the non-homogeneity in a progressive process with informative observations. Sweeting et al. (2010) developed a multi-state Markov model for disease progression in the presence of informative examinations by using a more regularly observed auxiliary variable. In this paper, we provide a general method to handle the non-homogeneity in Markov processes using the time-transformation method in the presence of informative observations. Our proposed method is more appealing than the method of Chen et al. (2010) since it requires to estimate fewer parameters, and hence can be employed when a smaller number of subjects or shorter observation periods are available; and our proposed method is more appealing than the method of Sweeting et al. (2010) because it can deal with all types of missing data mechanisms (including MAR and MNAR) without needing a more regularly observed auxiliary variable. Furthermore, using our non-parametric time-transformation model, we can accommodate temporal non-homogeneity without assuming that transition intensities follow any particular functional form of time. Thus, our proposed method is more flexible than previous methods dealing with non-

homogeneity. Maximum likelihood methods are used with parameter estimation, carried out via the EM algorithm (Dempster et al., 1977), and variance estimation is performed using the Louis' method (Louis, 1982).

The remainder of this paper is organized as follows. In Section 2, we briefly introduce the AD study. In Section 3, we review models and estimation methods for continuous time models. In Sections 4 and 5, we describe methods for the parameter estimation when data are MAR and MNAR, respectively. Empirical studies including the simulation studies and sensitivity analyses are implemented in Section 6. Data arising from an AD study are analyzed in Section 7. We conclude the paper with a general discussion in Section 8.

2 An Alzheimer's disease study

There are approximately 5 million people in the U.S. and more than 37 million people worldwide affected by AD. With AD, a person's memory and ability to learn and carry out daily activities such as talking and eating are gradually destroyed. As the disease progresses, individuals may also experience changes in personality and behavior. Unfortunately, there is no cure for AD and there is no way to predict how fast someone will progress through the stages of the disease. However, early Alzheimer's diagnosis and treatment can slow the progression of Alzheimer's symptoms. So, it is desirable to identify the risk factors that affect the transitions among the disease states such as normal cognition, MCI and dementia.

To study AD, 29 Alzheimer's Disease Centers (ADCs) have been funded by the National Institute on Aging (NIA), and the National Alzheimer's Coordinating Center (NACC) is responsible for developing and maintaining a database of participant information collected from these ADCs. A standardized clinical data set, called the Uniform Data Set (UDS), was determined and defined. The goal of the UDS is to provide ADC researchers a standard set of assessment procedures, collected longitudinally, to better characterize ADC participants with mild AD and mild cognitive impairment in comparison with non-demented controls (Beekly et al., 2007). The UDS was implemented from September 2005 as initial visits, and follow-up visits based on Clinical Task Force definitions were developed annually.

Some studies have found that amnesic MCI is transient because future evaluations could yield a reversion to normal cognition (here we group normal and "impaired, not MCI" and denote by normal cognition for simplicity) as opposed to progression to dementia. Our goal is to identify risk factors that affect the transitions among these states. In the literature of AD studies, time-dependent increases in the rate of progression of disease are a prominent feature. Yesavage et al. (2002) and Harezlak et al. (2003) studied the Markov process models using the piecewise constant transition intensities, and Slazar et al. (2007) addressed a discrete time Markov model for AD through the piecewise constant transition intensities. A disadvantage of this approach is that it requires a large study population to obtain precise estimates because it often requires estimation of more parameters. Additionally, transition rates must be assumed constant between any two cut points, which may not be desirable in some situations. Hubbard et al. (2008) considered a time-transformation method to address the non-homogeneity. Because of its use of continuous time model, this method allows the transition rates change continuously through time, which is more flexible in practice.

Another challenge of AD studies is the informative examinations. There are 7932 subjects from 29 ADCs included at the entry of this study. Follow-up visits for subjects are scheduled at approximately one-year intervals, with up to four clinical visits at present. During this study, patients may choose when they want to visit clinics for clinical examinations according to their degree of disease activity, for example. This creates the informative observations. There are 6722 subjects with complete data observed. The missing

proportion is 13.27%. Often, the easily implemented method such as the complete case analysis gives biased inferences.

In this paper, we address the informative examinations under the framework of the non-homogeneous process since little work has been devoted to this kind of study in literature. We use the time-transformation method (Hubbard et al., 2008) to address the non-homogeneity and use the EM algorithm to address the informative examinations when the observation times are pre-planned.

3 Notation and model formulation

3.1 Non-homogeneous Markov process model via time transformation

Suppose there are K states, $1, 2, \dots, K$, and let $Y(u)$ represent the state occupied at time $u \geq 0$, and $\mathcal{H}(u) = \{Y(v), 0 \leq v < u\}$ denote the history of the response process, which records the states occupied over the interval $[0, u)$. The transition probability function is generally written as $P\{Y(u+v) = k | Y(v) = \tilde{k}, \mathcal{H}(v)\}$ for $u, v > 0$, but under a Markov model this simplifies to $P\{Y(u+v) = k | Y(v) = \tilde{k}\}$, which we denote compactly as $P_{\tilde{k}k}(v, v+u)$. The Markov transition intensity function at time u for transitions from state $\tilde{k}$ to state k is given as follows:

$$q_{\tilde{k}k}^*(u) = \lim_{\Delta u \rightarrow 0} \frac{P\{Y(u+\Delta u) = k | Y(u) = \tilde{k}\}}{\Delta u}, \quad \tilde{k} \neq k,$$

$$q_{\tilde{k}k}^*(u) = -\sum_{k \neq \tilde{k}} q_{\tilde{k}k}^*(u), \quad \tilde{k} = k,$$

(Cox and Miller, 1977). A multi-state model with state space $\{1, 2, \dots, K\}$ can then be described via the transition intensity matrix $\mathbf{Q}^*(u)$ with elements $q_{\tilde{k}k}^*(u)$, $\tilde{k}, k = 1, \dots, K$.

To model the dependence of the transition intensities on risk factors, we may introduce covariates by expressing the transition intensities as functions of time (in the non-homogeneous case) and covariates. For a given individual i , we often adopt models of the form,

$$q_{i\tilde{k}k}^*(u) = q_{0\tilde{k}k}^*(u) g(\mathbf{X}_{i\tilde{k}k}; \beta_{\tilde{k}k}),$$

where $q_{i\tilde{k}k}^*(u)$ is the $\tilde{k}k$ -th element of the transition intensity matrix $\mathbf{Q}_i^*(u)$, $q_{0\tilde{k}k}^*(u)$ is the baseline transition intensity, $\mathbf{X}_{i\tilde{k}k}$ is the time-independent covariate vector, and $g(\cdot; \cdot)$ is any positive function. Let $\mathbf{X}_i = (\mathbf{X}_{i\tilde{k}k}', \tilde{k}, k = 1, \dots, K)'$.

Let $\mathbf{P}(v, v+u)$ denote the $K \times K$ transition probability matrix from time v to time $v+u$. For a homogeneous process we assume $q_{\tilde{k}k}^*(u) = q_{\tilde{k}k}^*$, and $P_{\tilde{k}k}(v, v+u) = P_{\tilde{k}k}(u)$ for $\tilde{k}, k = 1, \dots, K$. In the matrix form, we have

$$\mathbf{P}(u) = \exp(\mathbf{Q}^* u).$$

For a non-homogeneous process, we can perform some time scale transformation such that the process is homogeneous afterward (Hubbard et al., 2008). Specifically, let $t = h(u)$ be a time transformation on which the process is homogeneous with intensity matrix $\mathbf{Q}$, then

$$\mathbf{P}(u_1, u_2) = \mathbf{P}(t_2 - t_1) = \exp\{\mathbf{Q}(t_2 - t_1)\},$$

where $t_m = h(u_m)$, $m = 1, 2$. It is easy to show that $\mathbf{Q}^*(u) = \mathbf{Q}dh(u)/du$, which implies that time scale transformations leading to a time homogeneous Markov process are possible if the non-homogeneity in the process is due to a time-varying multiplicative change in the matrix of transition intensities.

Here we assume, after the time scale transformation $t = h(u)$, the transition intensity matrix for subject i does not depend on time. Then, the model becomes

$$q_{i\tilde{k}k}(t) = q_{0\tilde{k}k}g(\mathbf{X}_{i\tilde{k}k}; \beta_{\tilde{k}k}),$$

where $q_{i\tilde{k}k}$ is the $\tilde{k}k$ -th element of the homogeneous intensity matrix for subject i , $\mathbf{Q}_{\tilde{k}}$. Let $P_{i\tilde{k}k}(t)$ denote the transition probability for subject i from state $\tilde{k}$ to state k given the covariate vector $\mathbf{X}_{i\tilde{k}k}$.

The choice of $h(\cdot)$ is very flexible, and we require $h(u) \geq 0$ and $dh(u)/du \geq 0$, since $h(u)$ defines a time scale. Often, $h(\cdot)$ is selected by involving some parameter ϕ . The two examples used widely in practice, provided by Hubbard et al. (2008), are the exponential time transformation $h(u) = u\phi^u$ and the non-parametric time transformation $h(u) = u\phi(u)$, where

$$\begin{aligned}\varphi(u) &= \sum_{m=1}^d c(u)\varphi_m \left\{ \frac{1}{\gamma} K \left(\frac{u-u_m}{\gamma} \right) \right\}, \\ c(u) &= \left\{ \sum_{k=1}^d \frac{1}{\gamma} K \left(\frac{u-u_k}{\gamma} \right) \right\}^{-1}.\end{aligned}$$

This kernel smoother has knots at u_k , $k = 1, \dots, d$, smoothing parameter ϕ satisfies constraints $\phi_k > 0$. To make identifiability, we often assume $\phi_1 = 1$ or $\phi(0) = 0$.

We comment that not all non-homogeneous models can be so transformed to homogeneity; by selecting the transformation function, the proposed method can cover various non-homogeneity cases that are often used in practice. For example, using the non-parametric time-transformation model can accommodate temporal non-homogeneity without assuming that transformation intensities follow any particular functional form of time, which is more flexible than previous method dealing with non-homogeneity. The exponential transformation form has several advantages. First, it has a good interpretation: if $\phi > 1$, it means a $(\phi-1) \times 100\%$ increase in the rate of all transitions per year (assuming a yearly time unit); if $\phi < 1$, it means a $(1-\phi) \times 100\%$ decrease in the rate of all transitions per year (assuming a yearly time unit); $\phi = 1$ means the process is homogeneous on both the original and transformed time scales. Second, it requires estimation of fewer parameters, and hence can reduce the computation burdensome and can also be employed even a smaller number of subjects or shorter observation periods are available.

3.2 Maximum likelihood estimation under independent inspection process for complete data

With continuous time models and observation schemes, the response process $\{Y(u), u > 0\}$ may be observed at any time point u over the period observation. If the time of assessment u does not depend on the state of the underlying response process Y , we can base inference on

the response process conditional on the assessment times (Gruiger et al., 1991), and this is typically an implicit assumption in standard analyses. In this paper, we consider the problem in which subjects are scheduled to be examined at pre-specified assessment times denoted $u_1 < u_2 < \dots < u_J$. This reflects many common clinical settings where patients are expected to return for regular follow-up assessment on annual, say, basis. This enables us to adopt a convenient frame work employed to describe incomplete longitudinal data since it is then only necessary to indicate whether each assessment is made.

Let $\mathbf{Y}_i = \{Y_i(u_1), \dots, Y_i(u_J)\}'$ be a health state vector for subject i at the J observation time points, which takes values $1, \dots, K$, where $i=1, \dots, n$. Let β denote the parameter vector in the intensity matrix $\mathbf{Q}_i$, as well as the parameter ϕ in the time-transformation function $h(u, \phi)$. Using the time transformation and the homogeneity of the process on the new time scale, we can express the likelihood, conditional on the initial states, as follows:

$$L(\beta; \mathbf{Y}) = \prod_{i=1}^n \prod_{j=2}^J P[Y_i\{h(u_j; \phi)\} | Y_i\{h(u_{j-1}; \phi)\}, \mathbf{X}_i; \beta],$$

given the covariate vector $\mathbf{X}_i$. Maximizing $L(\beta; \mathbf{Y})$ with respect to β , we obtain the maximum likelihood estimate of β , say, $\hat{\beta}$.

4 Estimation procedure under the MAR mechanism

Here we focus on the likelihood-based method in which we need to specify the joint distribution of the observed response $\mathbf{Y}_i^{(o)}$ and the missing data indicator vector $\mathbf{R}_i$, given the covariate vector $\mathbf{X}_i$. Here $\mathbf{Y}_i$ is written as $\{Y_i^{(o)}, Y_i^{(m)}\}$ to indicate observed and missing components of the data for subject i , and $\mathbf{R}_i = \{R_i(u_1), \dots, R_i(u_J)\}'$ where $R_i(u_j) = 1$ if the response $Y_i(u_j)$ is observed, and 0 otherwise. The joint density of the observed data $(\mathbf{y}_i^{(o)}, \mathbf{r}_i)$ can be written as

$$f(\mathbf{r}_i, \mathbf{y}_i^{(o)} | \mathbf{X}_i; \alpha, \beta) = \int f(\mathbf{r}_i | \mathbf{y}_i^{(o)}, \mathbf{y}_i^{(m)}, \mathbf{X}_i; \alpha) f(\mathbf{y}_i^{(o)}, \mathbf{y}_i^{(m)} | \mathbf{X}_i; \beta) d\mathbf{y}_i^{(m)},$$

where α and β are the associated parameters for the missing data process and response process respectively. Then, the joint likelihood for (α, β) is given as follows:

$$L(\alpha, \beta; \mathbf{y}^{(o)}, \mathbf{r}) = \prod_{i=1}^n \int f(\mathbf{r}_i | \mathbf{y}_i^{(o)}, \mathbf{y}_i^{(m)}, \mathbf{X}_i; \alpha) f(\mathbf{y}_i^{(o)}, \mathbf{y}_i^{(m)} | \mathbf{X}_i; \beta) d\mathbf{y}_i^{(m)}. \quad (1)$$

When the missing data mechanism is missing completely at random (MCAR) or MAR, this likelihood becomes

$$\begin{aligned} L(\alpha, \beta; \mathbf{y}^{(o)}, \mathbf{r}) &= \prod_{i=1}^n \left\{ f(\mathbf{r}_i | \mathbf{y}_i^{(o)}, \mathbf{X}_i; \alpha) \int f(\mathbf{y}_i^{(o)}, \mathbf{y}_i^{(m)} | \mathbf{X}_i; \beta) d\mathbf{y}_i^{(m)} \right\} \\ &= \prod_{i=1}^n \{ f(\mathbf{r}_i | \mathbf{y}_i^{(o)}, \mathbf{X}_i; \alpha) f(\mathbf{y}_i^{(o)} | \mathbf{X}_i; \beta) \}. \end{aligned}$$

Assuming the parameters α and β are functionally independent, then likelihood-based inference for β from the likelihood $L(\alpha, \beta; \mathbf{y}^{(o)}, \mathbf{r})$ is equivalent to a likelihood-based inference for β from the following observed likelihood:

$$L(\beta; \mathbf{y}^{(o)}) = \prod_{i=1}^n f(\mathbf{y}_i^{(o)} | \mathbf{X}_i; \beta). \quad (2)$$

For the informative observations with MAR, we develop a Fisher-scoring method to obtain the maximum likelihood estimate in the following subsection.

4.1 A Fisher-scoring method

Let $j_1 < j_2 < \dots < j_{m_i}$ be the ordered observed assessment time points for subject i . We assume that data are always available in the baseline assessment, so $j_1 = 1$. The observed likelihood (2) can be written as follows:

$$\begin{aligned} L(\beta; \mathbf{y}^{(o)}) &= \prod_{i=1}^n \prod_{m=2}^{m_i} P[Y_i\{h(u_{j_m}; \varphi)\} | Y_i\{h(u_{j_{m-1}}; \varphi)\}, \mathbf{X}_i] \\ &= \prod_{i=1}^n \prod_{m=2}^{m_i} \prod_{k=1}^K [P_{i\tilde{k}k}\{h(u_{j_m}) - h(u_{j_{m-1}})\}]^{I[Y_i\{h(u_{j_{m-1}})\} = \tilde{k}, Y_i\{h(u_{j_m})\} = k]}, \end{aligned}$$

leading to the score function,

$$\mathbf{S}(\beta) = \sum_{i=1}^n \sum_{m=2}^{m_i} \sum_{k=1}^K \frac{I[Y_i\{h(u_{j_{m-1}})\} = \tilde{k}, Y_i\{h(u_{j_m})\} = k]}{P_{i\tilde{k}k}\{h(u_{j_m}) - h(u_{j_{m-1}})\}} \frac{\partial P_{i\tilde{k}k}\{h(u_{j_m}) - h(u_{j_{m-1}})\}}{\partial \beta},$$

and the second derivative

$$\begin{aligned} \frac{\partial^2 \log L(\beta; \mathbf{y}^{(o)})}{\partial \beta \partial \beta'} &= \sum_{i=1}^n \sum_{m=2}^{m_i} \sum_{k=1}^K \left\{ \frac{I[Y_i\{h(u_{j_{m-1}})\} = \tilde{k}, Y_i\{h(u_{j_m})\} = k]}{P_{i\tilde{k}k}\{h(u_{j_m}) - h(u_{j_{m-1}})\}} \frac{\partial^2 P_{i\tilde{k}k}\{h(u_{j_m}) - h(u_{j_{m-1}})\}}{\partial \beta \partial \beta'} \right. \\ &\quad \left. - \frac{I[Y_i\{h(u_{j_{m-1}})\} = \tilde{k}, Y_i\{h(u_{j_m})\} = k]}{[P_{i\tilde{k}k}\{h(u_{j_m}) - h(u_{j_{m-1}})\}]^2} \frac{\partial P_{i\tilde{k}k}\{h(u_{j_m}) - h(u_{j_{m-1}})\}}{\partial \beta} \frac{\partial P_{i\tilde{k}k}\{h(u_{j_m}) - h(u_{j_{m-1}})\}}{\partial \beta'} \right\}. \end{aligned}$$

To solve $\mathbf{S}(\beta) = 0$ in order to obtain the estimate $\hat{\beta}$, we may, in principle, apply the Newton–Raphson algorithm by using the observed information matrix. This requires the availability of the second derivatives of the log-likelihood. However, the second derivatives of the log-likelihood are tedious to derive and program. Here we develop a Fisher-scoring method, which obviates need for the second derivatives (Kalbfleisch and Lawless, 1985) when the covariates are discrete.

Taking the expectation with respect to the conditional distribution of the response vectors given the covariates, we obtain the following expectation:

$$E \left\{ -\frac{\partial^2 \log L(\beta; \mathbf{y}^{(o)})}{\partial \beta \partial \beta'} \right\} = \sum_{i=1}^n \sum_{m=2}^{m_i} \sum_{k=1}^K \left\{ \frac{P[Y_i\{h(u_{j_{m-1}})\} = \tilde{k} | \mathbf{X}_i]}{P_{i\tilde{k}k}\{h(u_{j_m}) - h(u_{j_{m-1}})\}} \cdot \frac{\partial P_{i\tilde{k}k}\{h(u_{j_m}) - h(u_{j_{m-1}})\}}{\partial \beta} \frac{\partial P_{i\tilde{k}k}\{h(u_{j_m}) - h(u_{j_{m-1}})\}}{\partial \beta'} \right\},$$

by noting that

$$E(I[Y_i\{h(u_{j_{m-1}})\}=\tilde{k}, Y_i\{h(u_{j_m})\}=k])=P[Y_i\{h(u_{j_{m-1}})\}=\tilde{k}|\mathbf{X}_i] \cdot P_{i\tilde{k}k}\{h(u_{j_m})-h(u_{j_{m-1}})\}$$

and

$$\sum_{k=1}^K P_{i\tilde{k}k}\{h(u_{j_m})-h(u_{j_{m-1}})\}=1.$$

This expectation can be estimated by

$$\mathbf{M}(\beta)=\sum_{i=1}^n \sum_{m=2}^{m_i} \sum_{k=1}^K \left\{ \frac{P_{ij_{m-1}\tilde{k}}}{P_{i\tilde{k}k}\{h(u_{j_m})-h(u_{j_{m-1}})\}} \cdot \frac{\partial P_{i\tilde{k}k}\{h(u_{j_m})-h(u_{j_{m-1}})\}}{\partial \beta} \cdot \frac{\partial P_{i\tilde{k}k}\{h(u_{j_m})-h(u_{j_{m-1}})\}}{\partial \beta'} \right\},$$

when covariates $\mathbf{X}_i$ are discrete, where $p_{ij_{m-1}\tilde{k}}$ is the proportion of the subjects with covariate $\mathbf{x}_i$ in state $\tilde{k}$ at the (j_{m-1}) th time point.

Then an updated estimate is obtained as

$$\beta^{(h+1)}=\beta^{(h)}+[\mathbf{M}(\beta^{(h)})]^{-1}\mathbf{S}(\beta^{(h)}), \quad h=0, 1, \dots,$$

where $\mathbf{M}(\beta^{(h)})$ is assumed non-singular. The iteration is cycled through until convergence of $\beta^{(h+1)}$. Let $\hat{\beta}$ denote the corresponding limit. Under the regularity conditions for maximum likelihood estimators, $\sqrt{n}(\hat{\beta}-\beta) \rightarrow N(0, \sum^{-1})$ as the sample size n approaches infinity. Here $\sum = E\{-\partial^2 \log L_i(\beta; \mathbf{y}_i^{(o)}) / \partial \beta \partial \beta'\}$ with $L_i(\beta; \mathbf{y}_i^{(o)})$ the observed likelihood for subject i . We can estimate the asymptotic covariance matrix of $\hat{\beta}$ by $[\mathbf{M}(\hat{\beta})]^{-1}$.

When the number K of states is relatively small, the Fisher-scoring method described above works well. However, if K is large, the Fisher-scoring method may become computationally burdensome. In the Appendix, we describe an alternative method – the EM algorithm, which is simple to implement.

5 Estimation procedure under the missing not at random mechanism

When data are MNAR, the missing data mechanism must be accommodated to make valid inference because the likelihood of the observed response $\mathbf{Y}^{(o)}$ and the missing indicator $\mathbf{R}$, given by (1), cannot be simplified to the observed-data likelihood. To model the joint distribution of $\mathbf{Y}^{(o)}$ and $\mathbf{R}$, we must characterize the missing data process and estimate the associated parameters. We address this in the following subsection.

5.1 Modeling for the missing data process

A multinomial missing data model, introduced by Ibrahim et al. (2001), is often used to specify the joint distribution of R_i through a sequence of one-dimensional conditional distributions:

$$P(\mathbf{R}_i | \mathbf{Y}_i, \mathbf{X}_i; \alpha) = \prod_{j=2}^J P\{R_i(u_j) | \tilde{\mathbf{R}}_i(u_j), \mathbf{Y}_i, \mathbf{X}_i; \alpha\} \cdot P\{R_i(u_1) | \mathbf{Y}_i, \mathbf{X}_i; \alpha\},$$

where $\tilde{\mathbf{R}}_i(u_j) = \{R_i(u_{j-1}), \dots, R_i(u_1)\}$, and α is a vector of associated parameters for the conditional distribution. This accommodates non-monotone patterns of missing data, and provides a natural way to specify the joint distribution of the missing data indicators when knowledge about the missingness of one response affects the probability of missingness of another. Let $\lambda_i(u_j) = P\{R_i(u_j) = 1 | \tilde{\mathbf{L}}_i(u_j); \alpha\}$ denote the conditional probability that the response is observed at time u_j given $\tilde{\mathbf{L}}_i(u_j) = \{\tilde{\mathbf{R}}_i(u_j), \mathbf{Y}_i, \mathbf{X}_i\}$. Typically a logistic link may relate a linear function of $\tilde{\mathbf{L}}_i(u_j)$ to the probability of being observed at time u_j for subject i , i.e. we can specify $\text{logit } \lambda_i(u_j) = \eta\{\tilde{\mathbf{L}}_i(u_j); \alpha\}$ for certain function $\eta\{\cdot; \cdot\}$. For example, we may consider the following model, although other forms may be well adopted for specific applications,

$$\text{logit } \lambda_i(u_j) = \alpha_0 + \alpha_1 r_i(u_{j-1}) + \alpha_2 r_i(u_{j-1}) y_i(u_{j-1}) + \alpha_3 y_i(u_j) + \alpha_4 (u_j - u_{j-1}) + \alpha'_x \mathbf{X}_i. \quad (3)$$

If $\alpha_3 = 0$, the missing mechanism specified by (3) leads to an MNAR mechanism; $\alpha_3 = 0$ but $\alpha_2 \neq 0$ leads to an MAR mechanism; both $\alpha_3 = 0$ and $\alpha_2 = 0$ leads to an MCAR mechanism. Let $\theta = (\beta', \alpha')'$.

5.2 An EM algorithm

Here we use the EM algorithm for parameter estimation under an MNAR mechanism. Conditional on the initial state, the complete data likelihood for subject i is written as

$$L_i(\theta; \mathbf{y}_i, \mathbf{r}_i) = \prod_{j=2}^J [\{\lambda_i(u_j)\}^{r_i(u_j)} \{1 - \lambda_i(u_j)\}^{1-r_i(u_j)} \times \prod_{\tilde{k}, k=1}^K [P_{\tilde{k}k}\{h(u_j) - h(u_{j-1})\}]^{I[Y_i\{h(u_{j-1})\}=\tilde{k}, Y_i\{h(u_j)\}=k]}],$$

leading to the log-likelihood

$$\ell_i(\theta; \mathbf{y}_i, \mathbf{r}_i) = \sum_{j=2}^J \left[r_i(u_j) \log\{\lambda_i(u_j)\} + \{1 - r_i(u_j)\} \log\{1 - \lambda_i(u_j)\} + \sum_{\tilde{k}, k=1}^K I[Y_i\{h(u_{j-1})\}=\tilde{k}, Y_i\{h(u_j)\}=k] \log[P_{\tilde{k}k}\{h(u_j) - h(u_{j-1})\}] \right].$$

In the E-step, we calculate the expectation of the complete data likelihood given the observed data and the h -th iteration parameter $\theta^{(h)}$. That is, $Q(\theta; \theta^{(h)}) = \sum_{i=1}^n Q_i(\theta; \theta^{(h)})$, where $Q_i(\theta; \theta^{(h)}) = E\{\ell_i(\theta; \mathbf{y}_i, \mathbf{r}_i) | \mathbf{y}_i^{(o)}; \theta^{(h)}\} = \sum_{\mathbf{y}_i^{(m)}} w_i(\mathbf{y}_i; \theta^{(h)}) \cdot \ell_i(\theta; \mathbf{y}_i, \mathbf{r}_i)$, and

$$w_i(\mathbf{y}_i; \theta^{(h)}) = \frac{L_i(\theta^{(h)}; \mathbf{y}_i^{(m)}, \mathbf{y}_i^{(o)}, \mathbf{r}_i)}{\sum_{\mathbf{y}_i^{(m)}} L_i(\theta^{(h)}; \mathbf{y}_i^{(m)}, \mathbf{y}_i^{(o)}, \mathbf{r}_i)}.$$

In the M-step, we maximize $Q(\theta; \theta^{(h)})$ to get the estimate $\theta^{(h+1)}$. Iterate E and M steps until convergence. Denote the limit $\hat{\theta}$. We comment that we can still use the Fisher-scoring method introduced in Section 4 to obtain the estimate when the covariate $\mathbf{X}_i$ is discrete, which we do not need to calculate the second derivatives of the transition probabilities.

Standard errors for these parameter estimates can be calculated using Louis' method (Louis, 1982), which partitions the complete data information into two parts: the information associated with the observed data and the one associated with the missing data. The estimated observed information matrix of θ based on Louis' method is given by

$$\sum(\hat{\theta}) = -\frac{\partial^2 Q(\hat{\theta}; \hat{\theta})}{\partial \theta \partial \theta'} - \sum_{i=1}^n \sum_{\mathbf{y}_i^{(m)}} w_i(\mathbf{y}_i; \hat{\theta}) \left\{ \frac{\partial \ell_i(\hat{\theta}; \mathbf{y}_i, \mathbf{r}_i)}{\partial \theta} \right\} \left\{ \frac{\partial \ell_i(\hat{\theta}; \mathbf{y}_i, \mathbf{r}_i)}{\partial \theta} \right\}' + \sum_{i=1}^n \left\{ \frac{\partial Q_i(\theta; \hat{\theta})}{\partial \theta} \right\} \left\{ \frac{\partial Q_i(\theta; \hat{\theta})}{\partial \theta} \right\}'.$$

The estimate of the asymptotic covariance matrix of $\hat{\beta}$ is the left upper $p \times p$ block of $[\Sigma(\hat{\theta})]^{-1}$, where p is the dimension of β .

6 Numerical studies

6.1 Model assessment for the proposed method

Here we consider a three-state transition process with transition intensity matrix

$$Q = \begin{pmatrix} -0.3 & 0.2 & 0.1 \\ 0.2 & -0.3 & 0.1 \\ 0 & 0 & 0 \end{pmatrix}.$$

We assume the time points are uniformly on (0.5, 2) with equal space interval 0.5. At the first observation time point, subjects are equally likely to be in state one or two. We assume the true-transformation function $h(u) = u\phi^u$ with true parameter $\phi = 1.2$.

The missing data model is

$$\text{logit } \lambda_i(u_j) = \alpha_0 + \alpha_1 r_i(u_{j-1}) + \alpha_2 r_i(u_{j-1}) y_i(u_{j-1}) + \alpha_3 y_i(u_j), \quad (4)$$

for $j = 2, 3, \dots$. The true values are $\alpha_0 = \log(0.4)$, $\alpha_1 = \log(1.2)$ and $\alpha_2 = \log(1.5)$, and we change α_3 to adjust the missing mechanism and missing proportions. One thousand simulations are run for each parameter configuration.

Here we compare two methods. One is the proposed method, and the second, entitled "complete case analysis," is the proposed method by removing subjects with missing visits where we assume data are MCAR. Table 1 reports the result for sample size 1000 (we also do simulation studies for a moderate sample size 200, and results are reported in the Supporting Information in the journal website), where BIAS is the relative bias; ASE is the

average standard error based on the Louis' formula for the proposed method and model based for the naive method; ESE is the empirical standard deviation; CP is the 95% coverage probability. It is seen that the proposed method gives satisfactory results with ignorable finite sample biases and good coverage probabilities. However, the complete case method yields large biases and poor coverage probabilities. It is also not surprising that the complete case analysis is substantially biased even for MAR ($\alpha_3 = 0$), since the complete case analysis is valid when data are MCAR.

6.2 Sensitivity analysis and model diagnosis

The validity of this approach relies on correct specification of the structure for the intensity, the time-transformation function, and correct modeling of the missing data process. Here we investigate the impact of model misspecification for the missing data process and time-transformation function $h(\cdot)$. The model setup for the transition intensity and transformation of time scale is the same as that in Section 6.1.

We first evaluate the asymptotic biases induced by misspecifying the link function in the missing data model with a correct specification for the time-transformation function. We assume the true missing data model is (4), while we use model

$$\text{probit } \lambda_i(u_j) = \alpha_0 + \alpha_1 r_i(u_{j-1}) + \alpha_2 r_i(u_{j-1}) y_i(u_{j-1}) + \alpha_3 y_i(u_j)$$

in the estimation procedure, where $\text{probit}(\cdot)$ is the cumulative distribution function of the standard normal distribution. Figure 1 contains plots of the asymptotic percent relative biases of the regression coefficients, for example, q_{12} , q_{13} and ϕ . It can be seen that as the absolute value of α_3 increases, the relative biases increase; however, the relative bias is very small, indicating that the estimates are less sensitive to the misspecification of the logit link to the probit link. It is not surprising that there is little difference between the two results, since the two distribution functions behave very similar; this result may not be necessarily imply that the estimation is not sensitive to misspecified link functions.

Next, we consider the impact of misspecifying the missing data process by assuming that it is MAR when in fact it is MNAR, with a correct specification for the time-transformation function. We assume the true missing data model is (4), while we use model

$$\text{logit } \lambda_i(u_j) = \alpha_0^* + \alpha_1^* r_i(u_{j-1}) + \alpha_2^* r_i(u_{j-1}) y_i(u_{j-1})$$

in the estimation procedure. Figure 2 contains plots of the asymptotic percent relative biases of the regression coefficients q_{12} , q_{13} and ϕ . It can be seen that as the absolute value of α_3 increases, the relative biases increase; the estimated transition intensity parameters are more sensitive to misspecification of the missing data models, while the estimated transformation parameter $\hat{\phi}$ is less sensitive.

Last we consider the impact of misspecifying the time-transformation function but assuming a correct specification for the missing data model (4). Specifically, we use the time-transformation model $h(u) = u$, while the true model is $h(u) = u\phi^u$. Figure 3 contains plots of the asymptotic percent relative biases of the regression coefficients q_{12} and q_{13} . It can be seen that as the value of ϕ goes far away from 1, the relative biases increase.

As demonstrated above, estimation of β is sensitive to misspecification of the missing data and time-transformation models. Therefore, careful assessment of these models is warranted.

In general, a likelihood ratio test is used to compare the fit of two models, one of which is nested within the other. This often occurs when testing whether a simplifying assumption for a model is valid, as when two or more model parameters are assumed to be related. Both models are fitted to the data and their log-likelihood recorded. The test statistic is twice the difference in these log-likelihoods. In many cases, the probability distribution of the test statistic can be approximated by a chi-square distribution with k degrees of freedom, where k is the difference of the number of parameters between the full model and the reduced model. The model with more parameters will always fit at least as well (have a greater log-likelihood). Whether it fits significantly better and should thus be preferred can be determined by deriving the p -value of the obtained test statistic. For non-nested models, people may consider Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), etc. Alternatively, Gentleman et al. (1994) and Saint-Pierre et al. (2003) discuss the use of empirical and predicted state occupancy to assess goodness-of-fit for Markov process models. The idea is that we compare the observed and predicted prevalences of states at each time point, which would allow us to assess if the time-transformation assumption or the missing data model is reasonable.

In many applications, however, data may not be sufficient to distinguish between alternative non-ignorable missing data mechanisms. For this reason, sensitivity analyses are often advocated (e.g. Ibrahim et al., 2005). Several authors have proposed the use of global and local influence tools to sensitivity analyses in missing data contexts (e.g. Verbeke et al., 2001; Verbeke and Molenberghs, 2000; Molenberghs et al., 2001; Molenberghs and Verbeke, 2005; Zhu and Lee, 2001; Van Steen et al., 2001; Kenward, 1998; Jansen et al., 2003). Another route for sensitivity analysis is to consider pattern-mixture models as a complement to selection models. Other approaches have been considered by Copas and Eguchi (2001), Copas and Li (1997) and Copas and Shi (2000). More detail on some of these procedures can be found in Molenberghs and Verbeke (2005), and Little and Rubin (2002).

7 Application to the Alzheimer's disease study

In this section, we apply our proposed method to the AD study, described in Section 2, to investigate factors for transitions among normal cognition, MCI, dementia and death, where we treat death as an absorbing state but allowing transitions among all other states. Table 2 lists the total number of transitions between clinical diagnosis states at successive visits. Risk factor vector $\mathbf{X}_i$ includes: sex, congestive heart failure (CVCHF, yes/no), geriatric depression score (GDS), family history of dementia (fhdem, yes/no), diabetes (yes/no), hypertension (yes/no), education (years), MMSE and age at the time origin. Table 3 lists the baseline risk factors for the 7932 subjects.

For simplicity, the four states, normal cognition, MCI, dementia and death are coded as 1, 2, 3 and 4, respectively. The multiplicative models for transition $\tilde{k}$ to k after the transformation are

$$q_{i\tilde{k}k}(t) = q_{0\tilde{k}k} \exp(\mathbf{X}_i' \beta_{\tilde{k}k}),$$

for $\tilde{k}, k = 1, 2, 3, 4, \tilde{k} \neq k$.

The missing data model is

$$\text{logit } \lambda_i(u_j) = \alpha_0 + \mathbf{X}_i' \alpha_x + \alpha_1 I\{Y_i(u_j)=2\} + \alpha_2 I\{Y_i(u_j)=3\}.$$

Here, we compare two methods in line with the simulation study: the complete case analysis by assuming data are MCAR and the proposed method. In this study we assume an exponential transformation form $h(u) = u\phi^u$. The estimates of ϕ are 1.033 with 95% confidence interval (1.015 and 1.051) for the complete case analysis and 1.046 with 95% confidence interval (1.030 and 1.062) for the proposed method. Both methods reveal that the process exhibits significant non-homogeneity, and the rate of evolution of the process is increasing as a function of time. Figures 4 and 5 illustrate the baseline intensities from normal cognition to MCI and MCI to dementia, respectively. For model selections of the transition and missing data models, the likelihood ratio test discussed in Section 6.2 is employed. Final results for the transition and missing data models are reported in Tables 4 and 5, respectively. To investigate the goodness-of-fit for our time-transformation model, we compare the expected and observed state occupancies, as discussed in Section 6.2. The observed and expected state occupancies at each clinical visits are presented in Table 6. No significant difference is observed between the observed and expected state occupancies, indicating that the model fits for the transition intensities, the missing data and the time transformation are reasonable here.

Table 4 lists risk factors in the transitions from normal to MCI, MCI to dementia and dementia to death. Complete results are reported in the Supporting Information in the journal website. The complete case analysis and the proposed method give different estimates. In the transitions from normal cognition to MCI, family history of dementia and age are significant, indicating that older people and people who have a family history of dementia have higher risk of transition from normal cognition to MCI. In the transitions from MCI to dementia, sex, GDS and age are significant, indicating that older people or people with higher GDS score have higher risk of transition from MCI to dementia, and women have lower risk of transition from MCI to dementia. In the transition from dementia to death, sex, CVCHF, GDS, education and age are significant, indicating that a person has higher risk to death if he/she has a congestive heart failure, higher GDS score, higher education or with an older age, and women has lower risk from dementia to death comparing to men.

For the missing data model, significance of sex indicates that women are more likely to miss the observation than men; significance of fhdem and hypertension indicates that people with family history of dementia are less likely to attend this study, and people with hypertension are more likely to attend this study; significance of MMSE indicates that people with higher MMSE score are less likely to attend this study; significance of present state occupation indicate that the MNAR mechanism is perhaps reasonable, and people are less likely to attend the study if they are in MCI or dementia comparing to normal cognition.

8 Discussion

In this paper, we propose a likelihood-based method for the analysis of incomplete observations under the framework of non-homogeneous Markov processes using the time-transformation model. This method is very appealing in that it can deal with all kinds of missing data mechanisms and allow variation in transition intensities. Several methods are introduced for estimations with informative examinations under MAR or MNAR, including the Fisher-scoring method and the EM algorithm. The Fisher-scoring method is appealing since we do not need the second derivatives of the likelihood if covariates are discrete.

We apply the proposed method to the NACC UDS. It reveals that there is a significant increase in transition intensities as a function of time, which is non-homogeneous. Furthermore, we identified that the MAR assumption may not be feasible, and hence the method ignoring the missing data gives invalid inference. For example, in the transition from normal to dementia, the complete case analysis indicates that family history of dementia has a significant effect on this transition, but the proposed method reveals that family history of dementia is not significant. This difference may be caused by bias inherent in the complete case analysis. Perhaps the dependence of transition from normal to dementia on the family history of dementia is more marked for people who are observed.

Note that, to obtain consistent parameter estimates under MNAR mechanisms, both the transition and missing data models must be correctly specified. In practice, we aim to build a model which provides useful insight into the response process and observation process. Our strategy is therefore to build models that contain a large number of covariates, carry out tests of fit of nested models, and ultimately find a parsimonious model using standard procedures for model selection. The need for generalizations to deal with more complex models can be assessed by model expansion and the use of general model selection procedures via likelihood ratio tests.

Our method here assumes that the covariates are time independent, which is a limitation as most of the other methods to deal with transition models. Time-dependent covariates are not a problem if they are piecewise-constant, more problematic if they are known at all time points but continuously changing (e.g. age), and very problematic if they are only known at their observation times (e.g. biomarkers). Relatively little work has been done on fitting regression models with time-dependent covariates. In the special case of a single interval-censored covariate that indicates the development of a particular condition, Goggins et al. (1999) developed methods for Cox regression for a right censored event time. Chen and Cook (2003) considered models and methods to deal with an interval-censored progressive covariate process in recurrent event analyses. Cook et al. (2008) considered an extension to the bivariate setting where both the covariate and failure times are interval-censored. The more general problem of interval-censored time-varying covariates remains relatively open and worthy of future research.

Acknowledgments

The authors thank the Editor, Associated Editor and two referees for their constructive comments and suggestions. Dr. Xiao-Hua Zhou, Ph.D., is presently a Core Investigator and Biostatistics Unit Director at the Northwest HSR&D Center of Excellence, Department of Veterans Affairs Medical Center, Seattle, WA. Dr. Zhou's work was supported in part by U.S. Department of Veterans Affairs, Veterans Affairs Health Administration, HSR&D grants, and the National Science Foundation of China (NSFC 30728019). Both Drs. Zhou and Chen were supported in part by NIA grant U01AG016976. This paper presents the findings and conclusions of the authors. It does not necessarily represent those of VA HSR&D Service.

References

- Albert PS, Waclawiw MA. A two-state Markov chain for heterogeneous transitional data: a quasi-likelihood approach. *Statistics in Medicine*. 1998; 17:1481–1493. [PubMed: 9695193]
- Andersen, PK.; Borgan, O.; Gill, RD.; Keiding, N. *Statistical Models Based on Counting Processes*. Springer; New York, NY: 1993.
- Bartholomew DJ. Some recent developments in social statistics. *International Statistical Review*. 1983; 51:1–9.
- Beekly DL, Ramos EM, Lee WW, Deitrich WD, Jacka ME, Wu J, Hubbard JL, Koepsell TD, Morris JC, Kukull WA. the NIA ADCs. The National Alzheimer's Coordinating Center (NACC) database: the Uniform Data Set. *Alzheimer Disease and Related Disorders*. 2007; 21:249–258.

- Breto C, He D, Ionides EL, King AA. Time series analysis via mechanistic models. *Annals of Applied Statistics*. 2009; 3:319–348.
- Cappe, O.; Moulines, E.; Ryden, T. *Inference in Hidden Markov Models*. Springer Verlag; New York: 2005.
- Chen EB, Cook RJ. Regression modeling with recurrent events and time-dependent interval-censored marker data. *Lifetime Data Analysis*. 2003; 9:275–291. [PubMed: 14649846]
- Chen B, Yi GY, Cook RJ. Analysis of interval-censored disease progression data via multi-state models under a nonignorable inspection process. *Statistics in Medicine*. 2010; 29:1175–1189. [PubMed: 20437455]
- Cook RJ, Kalbfleisch JD, Yi GY. A generalized mover-stayer model for panel data. *Biostatistics*. 2002; 3:407–420. [PubMed: 12933606]
- Cook RJ, Yi GY, Lee KA, Gladman DD. A conditional Markov model for clustered progressive multistate processes under incomplete observation. *Biometrics*. 2004; 60:436–443. [PubMed: 15180669]
- Cook RJ, Zeng L, Lee KA. A multistate model for bivariate interval-censored failure time data. *Biometrics*. 2008; 64:1100–1109. [PubMed: 18218064]
- Copas JB, Eguchi S. Local sensitivity approximations for selectivity bias. *Journal of the Royal Statistical Society, Series B*. 2001; 63:871–895.
- Copas JB, Li HG. Inference for non-random samples (with discussion). *Journal of the Royal Statistical Society, Series B*. 1997; 59:55–96.
- Copas JB, Shi JQ. Meta-analysis, funnel plots and sensitivity analysis. *Biometrics*. 2000; 1:247–262.
- Cox, DR.; Miller, HD. *The Theory of Stochastic Processes*. Chapman and Hall; London: 1977.
- Dempster AP, Laird NM, Rubins DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*. 1977; 39:1–38.
- Gentleman RC, Lawless JF, Lindsey JC, Yan P. Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine*. 1994; 13:805–821. [PubMed: 7914028]
- Goggins WB, Finkelstein DM, Zaslavsky AM. Applying the Cox proportional hazards model when the change time of a binary time-varying covariate is interval-censored. *Biometrics*. 1999; 55:445–451. [PubMed: 11318198]
- Grüger J, Kay R, Schumacher M. The validity of inferences based on incomplete observations in disease state models. *Biometrics*. 1991; 47:595–605. [PubMed: 1912263]
- Harezlak J, Gao S, Hui S. An illness-death stochastic model in the analysis of longitudinal dementia data. *Statistics in Medicine*. 2003; 22:1465–1475. [PubMed: 12704610]
- Hubbard RA, Inoue LYT, Fann JR. Modeling nonhomogeneous Markov processes via time transformation. *Biometrics*. 2008; 64:843–850. [PubMed: 18047532]
- Ibrahim JG, Chen MH, Lipsitz SR. Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. *Biometrika*. 2001; 88:551–564.
- Ibrahim JG, Chen MH, Lipsitz SR, Herring AH. Missing-data methods for generalized linear models: a comparative review. *Journal of the American Statistical Association*. 2005; 100:332–346.
- Jansen I, Molenberghs G, Aerts M, Thjis H, Van Steen K. A local influence approach to binary data from a psychiatric study. *Biometrics*. 2003; 59:410–419. [PubMed: 12926726]
- Kalbfleisch JD, Lawless JF. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*. 1985; 80:863–871.
- Kalbfleisch, JD.; Lawless, JF. *Proceedings of the Statistics Canada Symposium on Analysis of Data in Time*. Ottawa, Ontario: Statistics Canada; 1989. Some statistical methods for panel life history data; p. 185–192.
- Kenward MG. Selection models for repeated measurements with nonrandom dropout: an illustration of sensitivity. *Statistics in Medicine*. 1998; 17:2723–2732. [PubMed: 9881418]
- Laird NM. Missing data in longitudinal studies. *Statistics in Medicine*. 1988; 7:305–315. [PubMed: 3353609]
- Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2. Wiley; New York: 2002.

- Louis T. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B.* 1982; 44:226–233.
- Molenberghs G, Kenward MG, Goetghebeur E. Sensitivity analysis for incomplete contingency tables. *Applied Statistics.* 2001; 50:15–29.
- Molenberghs, G.; Verbeke, G. *Models for Discrete Longitudinal Data.* Springer; New York, NY: 2005.
- Ocana-Riola R. Non-homogeneous Markov processes for biomedical data analysis. *Biometrical Journal.* 2005; 47:369–376. [PubMed: 16053260]
- Perez-Ocon R, Ruiz-Castro JE, Gamiz-Perez ML. Non-homogeneous Markov models in the analysis of survival after breast cancer. *Journal of the Royal Statistical Society, Series C.* 2001; 50:111–124.
- Saint-Pierre P, Combescure C, Daures JP, Godard P. The analysis of asthma control under a Markov assumption with use of covariates. *Statistics in Medicine.* 2003; 22:3755–3770. [PubMed: 14673936]
- Slazar J, Schmitt F, Yu L, Mendiondo M, Kryscio R. Shared random effects analysis of multi-state Markov models: application to a longitudinal study of transitions to dementia. *Statistics in Medicine.* 2007; 26:568–580. [PubMed: 16345024]
- Singer B, Spilerman S. The representation of social processes by Markov models. *American Journal of Sociology.* 1976a; 82:1–54.
- Singer B, Spilerman S. Some methodoloical issues in the analysis of longitudinal surveys. *Annals of Economic and Sociological Measurement.* 1976b; 5:447–474.
- Sweeting MJ, Farewell VT, De Angelis D. Multi-state Markov models for disease progression in the presence of informative examination times: an application to hepatitis C. *Statistics in Medicine.* 2010; 29:1161–1174. [PubMed: 20437454]
- Van Steen K, Molenberghs G, Thijs H. A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. *Statistical Modelling: An International Journal.* 2001; 1:125–142.
- Verbeke, G.; Molenberghs, G. *Linear Mixed Models for Longitudinal Data.* Springer; New York, NY: 2000.
- Verbeke G, Molenberghs G, Thijs H, Lesaffre E, Kenward MG. Sensitivity analysis for nonrandom dropout: a local influence approach. *Biometrics.* 2001; 57:43–50. [PubMed: 11252617]
- Wasserman S. Analyzing social networks as stochastic processes. *Journal of the American Statistical Association.* 1980; 75:280–294.
- Yesavage J, O'Hara R, Kraemer H, Noda A, Taylor J, Ferris S, Gely-Nargeot M, Rosen A, Friedman L, Sheikh J, Derouesne C. Modeling the prevalence and incidence of Alzheimer's disease and mild cognitive impairment. *Journal of Psychiatric Research.* 2002; 36:281–286. [PubMed: 12127595]
- Zhu HT, Lee SY. Local inference for incomplete data models. *Journal of the Royal Statistical Society, Series B.* 2001; 63:111–126.

Appendix: An EM algorithm for the MAR mechanism

Note that the complete data likelihood for subject i is given by

$$L_i(\beta; \mathbf{y}_i) = \prod_{j=2}^J \prod_{k=1}^K [P_{ikk}\{h(u_j) - h(u_{j-1})\}]^{I[Y_i\{h(u_{j-1})\}=\tilde{k}, Y_i\{h(u_j)\}=k]},$$

leading to the complete data log-likelihood

$$\ell_i(\beta; \mathbf{y}_i) = \sum_{j=2}^J \sum_{k=1}^K I[Y_i\{h(u_{j-1})\}=\tilde{k}, Y_i\{h(u_j)\}=k] \log[P_{ikk}\{h(u_j) - h(u_{j-1})\}].$$

In the expectation step (E-step), with the estimate $\beta^{(h)}$ at iteration h , we construct the conditional expectation $Q(\beta; \beta^{(h)}) = \sum_{i=1}^n Q_i(\beta; \beta^{(h)})$, where $Q_i(\beta; \beta^{(h)}) = E\{\ell_i(\beta; \mathbf{y}_i) | \mathbf{y}_i^{(o)}, \beta^{(h)}\} = \sum_{\mathbf{y}_i^{(m)}} w_i(\mathbf{y}_i; \beta^{(h)}) \cdot \ell_i(\beta; \mathbf{y}_i)$, and

$$w_i(\mathbf{y}_i; \beta^{(h)}) = \frac{L_i(\beta^{(h)}; \mathbf{y}_i^{(o)}, \mathbf{y}_i^{(m)})}{\sum_{\mathbf{y}_i^{(m)}} L_i(\beta^{(h)}; \mathbf{y}_i^{(o)}, \mathbf{y}_i^{(m)})}.$$

The maximization step (M-step) maximizes the function $Q(\beta; \beta^{(h)})$ with respect to the parameter β , and a Newton–Raphson algorithm can be used for this purpose. Alternatively, we may employ the Fisher-scoring method discussed in Section 4.1. Specifically, the score and the expectation of the second derivative are given by

$$\mathbf{S}(\beta; \beta^{(h)}) = \sum_{i=1}^n \sum_{\mathbf{y}_i^{(m)}} w_i(\mathbf{y}_i; \beta^{(h)}) \cdot \partial \ell_i(\beta; \mathbf{y}_i) / \partial \beta,$$

and

$$\mathbf{M}(\beta; \beta^{(h)}) = E \left\{ -\frac{\partial^2 Q(\beta; \beta^{(h)})}{\partial \beta \partial \beta'} \right\} = \sum_{i=1}^n \sum_{\mathbf{y}_i^{(m)}} w_i(\mathbf{y}_i; \beta^{(h)}) \cdot E \{ -\partial^2 \ell_i(\beta; \mathbf{y}_i) / \partial \beta \partial \beta' \},$$

where

$$\frac{\partial \ell_i(\beta; \mathbf{y}_i)}{\partial \beta} = \sum_{j=2}^J \sum_{\tilde{k}, k=1}^K \frac{I[Y_i\{h(u_{j-1})\} = \tilde{k}, Y_i\{h(u_j)\} = k]}{P_{i\tilde{k}k}\{h(u_j) - h(u_{j-1})\}} \frac{\partial P_{i\tilde{k}k}\{h(u_j) - h(u_{j-1})\}}{\partial \beta},$$

and

$$E \left\{ -\frac{\partial^2 \ell_i(\beta; \mathbf{y}_i)}{\partial \beta \partial \beta'} \right\} = \sum_{j=2}^J \sum_{\tilde{k}, k=1}^K \left\{ \frac{P[Y_i\{h(u_{j-1})\} = \tilde{k} | \mathbf{X}_i]}{P_{i\tilde{k}k}\{h(u_j) - h(u_{j-1})\}} \frac{\partial P_{i\tilde{k}k}\{h(u_j) - h(u_{j-1})\}}{\partial \beta} \cdot \frac{\partial P_{i\tilde{k}k}\{h(u_j) - h(u_{j-1})\}}{\partial \beta'} \right\}.$$

If the covariate vector $\mathbf{X}_i$ are discrete, we can replace $P[Y_i\{h(u_{j-1})\} = \tilde{k} | \mathbf{X}_i]$ by the proportion of subjects that are in state $\tilde{k}$ at time point u_{j-1} with covariate $\mathbf{X}_i$.

To obtain the variance estimates for $\hat{\beta}$, we may apply Louis' method (Louis, 1982):

$$\sum(\hat{\beta}) = \mathbf{M}(\hat{\beta}; \hat{\beta}) - \sum_{i=1}^n \sum_{\mathbf{y}_i^{(m)}} w_i(Y_i; \hat{\beta}) \left\{ \frac{\partial \ell_i(\hat{\beta})}{\partial \beta} \right\} \left\{ \frac{\partial \ell_i(\hat{\beta})}{\partial \beta} \right\}' + \sum_{i=1}^n \left\{ \frac{\partial Q_i(\hat{\beta}; \hat{\beta})}{\partial \beta} \right\} \left\{ \frac{\partial Q_i(\hat{\beta}; \hat{\beta})}{\partial \beta} \right\}',$$

then $[\Sigma(\hat{\beta})]^{-1}$ is the estimate of the asymptotic covariance matrix of $\hat{\beta}$.

Figure 1.

Asymptotic percent relative biases of transition intensity with correctly specified linear predictor but the link is misspecified as probit (true link is logit): q_{12} and q_{13} are the transition intensities from state 1 to state 2 and 3 respectively, and ϕ is the time-transformation parameter.

Figure 2.

Asymptotic percent relative biases of transition intensity with correctly specified link but the linear predictor is misspecified ($\alpha_3 \neq 0$ so data are MNAR but are assumed MAR): q_{12} and q_{13} are the transition intensities from state 1 to state 2 and 3 respectively, and ϕ is the time-transformation parameter.

Figure 3.

Asymptotic percent relative biases of transition intensity with correctly specified missing data model but the time-transformation model is misspecified: q_{12} and q_{13} are the transition intensities from state 1 to state 2 and 3 respectively.

Figure 4.
Baseline transition intensities from Normal to MCI.

Figure 5.
Baseline transition intensities from MCI to Dementia.

Empirical performance of the proposed method and naive method for simulation studies with sample size $n = 1000$: the missing proportions are about 13, 25 and 40% with $\alpha_3 = \log(5.0)$, $\log(2.0)$ and $\log(1.0)$.

Table 1

α_3	Parameters	Proposed method				Complete case			
		BIAS%	ASE	ESE	CP%	BIAS%	ASE	ESE	CP%
$\log(5.0)$	q_{12}	-0.0	0.036	0.035	94.2	97.7	0.041	0.051	01.9
	q_{13}	-0.6	0.018	0.018	94.8	109.5	0.021	0.025	00.6
	q_{21}	1.2	0.062	0.061	94.2	37.0	0.075	0.097	73.1
	q_{23}	-1.0	0.037	0.037	95.0	60.1	0.042	0.049	63.4
$\log(2.0)$	ϕ	0.9	0.068	0.067	95.2	-19.3	0.080	0.104	38.0
	q_{12}	-1.0	0.030	0.030	95.1	171.1	0.105	0.106	10.6
	q_{13}	-0.8	0.015	0.014	95.4	139.4	0.046	0.046	11.4
	q_{21}	-1.0	0.050	0.050	94.7	124.4	0.184	0.191	42.5
$\log(1.0)$	q_{23}	-0.5	0.032	0.032	95.0	81.1	0.095	0.100	23.5
	ϕ	0.1	0.067	0.066	94.5	-61.2	0.238	0.242	58.5
	q_{12}	-1.0	0.035	0.034	95.4	113.0	0.402	0.405	13.8
	q_{13}	-1.0	0.018	0.017	95.6	234.9	0.143	0.139	20.6
	q_{21}	-0.5	0.060	0.060	95.0	138.9	0.534	0.540	19.2
	q_{23}	-1.6	0.038	0.039	94.5	95.8	0.292	0.291	26.9
	ϕ	0.4	0.067	0.064	95.6	-43.1	0.422	0.429	34.6

Table 2

Number of transitions made between states at successive clinic visits.

	?	Normal	MCI	Dementia	Death
?	250 (15.6%)	597 (37.2%)	247 (15.4%)	377 (23.5%)	132 (8.2%)
Normal	1150 (20.8%)	3865 (69.8%)	334 (6.0%)	57 (1.0%)	129 (2.3%)
MCI	505 (21.5%)	244 (10.4%)	1162 (49.5%)	343 (14.6%)	94 (4.0%)
Dementia	776 (21.3%)	35 (1.0%)	54 (1.5%)	2387 (65.6%)	386 (10.6%)

Table 3

Baseline risk factors for the 7962 subjects: proportions for categorical covariates and mean (standard deviation) for continuous covariates.

Risk factor	Summary
Female (%)	55.9
Congestive heart failure (%)	4.0
GDS	2.0 (2.5)
Family history of dementia (%)	27.9
Diabetes (%)	12.6
Hypertension (%)	54.5
Years of education	15.7 (8.5)
MMSE	26.3 (8.1)
Age	75.5 (9.3)

Table 4

Comparisons of two methods for the multiplicative effects on the transition intensities in the studies of Alzheimer's disease: hazard ratios and 95% confidence intervals.

Parameter	Complete case			Proposed method		
	HR	95%LCL	95%UCL	HR	95%LCL	95%UCL
Normal→MCI						
SEX (F)	1.06	0.47	2.38	1.02	0.83	1.24
CVCHF	1.00	0.58	1.74	1.00	0.60	1.67
GDS	0.99	0.76	1.31	0.99	0.73	1.34
fhdem	1.39	1.11	1.74	1.27	1.03	1.57
diabete	1.01	0.20	5.17	1.20	0.89	1.62
hypert	0.97	0.37	2.55	0.93	0.27	3.15
EDUC	1.01	0.99	1.03	1.01	0.97	1.05
MMSE	1.02	0.97	1.08	1.03	0.95	1.11
AGE	1.03	1.02	1.04	1.02	1.01	1.04
MCI→Dementia						
SEX (F)	0.71	0.59	0.85	0.71	0.60	0.85
CVCHF	0.54	0.27	1.11	0.52	0.26	1.00
GDS	1.08	1.05	1.13	1.08	1.04	1.12
fhdem	1.13	0.93	1.38	1.18	0.98	1.41
diabete	1.00	0.75	1.34	1.03	0.78	1.36
hypert	1.01	0.84	1.22	1.07	0.90	1.27
EDUC	0.99	0.96	1.01	0.98	0.97	1.00
MMSE	1.00	0.98	1.02	1.00	0.98	1.02
AGE	1.02	1.01	1.03	1.02	1.01	1.04
Dementia→Death						
SEX	0.67	0.63	0.72	0.65	0.59	0.71
CVCHF	1.74	1.57	1.91	1.78	1.55	2.04
GDS	1.14	1.13	1.15	1.13	1.12	1.14
fhdem	1.06	0.97	1.15	1.12	0.93	1.36
diabete	0.82	0.59	1.16	0.81	0.60	1.09
hypert	1.01	0.94	1.09	0.97	0.88	1.07

Parameter	Complete case			Proposed method		
	HR	95%LCL	95%UCL	HR	95%LCL	95%UCL
EDUC	1.01	1.00	1.02	1.01	1.00	1.02
MMSE	0.88	0.88	0.89	0.88	0.87	0.88
AGE	1.05	1.05	1.05	1.05	1.04	1.05

Table 5

Missing data model in the analysis of Alzheimer's disease.

Parameter	Estimate	SE	p-value
Intercept	3.17	0.27	<0.01
SEX (F)	-0.23	0.06	<0.01
CVCHF	0.16	0.15	0.29
GDS	-0.01	0.01	0.69
fhdem	-1.36	0.06	<0.01
diabete	-0.15	0.08	0.08
hypert	0.16	0.06	0.01
EDUC	0.01	0.00	0.21
MMSE	-0.01	0.00	0.01
AGE	-0.00	0.00	0.20
I (MCI)	-1.21	0.07	<0.01
I (Dementia)	-0.68	0.07	<0.01

Table 6

Observed and expected state occupancies at each clinic visit.

Visit	State occupancies (observed/expected)			
	Normal	MCI	Dementia	Death
1	0.43/0.43	0.21/0.21	0.36/0.36	0.00/0.00
2	0.47/0.47	0.20/0.20	0.31/0.31	0.02/0.02
3	0.45/0.45	0.19/0.19	0.31/0.31	0.05/0.05
4	0.45/0.45	0.16/0.16	0.31/0.31	0.07/0.07