



# A scalable approach for continuous time Markov models with covariates

FARHAD HATAMI

*Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield, Department of Medicine, University of Oxford and Department of Statistics, University of Oxford, Oxford, OX3 7LF, UK*

ALEX OCAMPO, GORDON GRAHAM

*Novartis Pharma AG, CH-4056 Basel, Switzerland*

THOMAS E. NICHOLS\* 

*Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford, OX3 7LF, UK and Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, OX3 9DU, UK*  
thomas.nichols@bdi.ox.ac.uk

HABIB GANJGAHI

*Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield, Department of Medicine, University of Oxford and Department of Statistics, University of Oxford, Oxford, OX3 7LF, UK and Department of Statistics, University of Oxford, 24-29 St Giles', Oxford, OX1 3LB, UK*

## SUMMARY

Existing methods for fitting continuous time Markov models (CTMM) in the presence of covariates suffer from scalability issues due to high computational cost of matrix exponentials calculated for each observation. In this article, we propose an optimization technique for CTMM which uses a stochastic gradient descent algorithm combined with differentiation of the matrix exponential using a Padé approximation. This approach makes fitting large scale data feasible. We present two methods for computing standard errors, one novel approach using the Padé expansion and the other using power series expansion of the matrix exponential. Through simulations, we find improved performance relative to existing CTMM methods, and we demonstrate the method on the large-scale multiple sclerosis NO.MS data set.

**Keywords:** Continuous-time Markov model; Multiple sclerosis; Multistate model; Padé approximation; Scalable optimization.

\*To whom correspondence should be addressed.

## 1. INTRODUCTION

Multiple sclerosis (MS) is a chronic neuroinflammatory and neurodegenerative disease that affects the central nervous system. Progression of the disease over time causes increasing disability, which is measured on the expanded disability status scale (EDSS) (Kurtzke, 1983). The EDSS measures several aspects of disability (such as walking) and reflects the functioning ability of an individual. The higher the EDSS score, the more disabled the individual. For instance, an EDSS score of 0 reflects a normal neurological examination, EDSS 6 indicates the use of a walking cane, and EDSS 10 is death. As part of clinical trials, the EDSS score is an ordinal outcome measured approximately every 4–6 months as a part of clinical trials in relapsing and progressive MS patients. To understand better how the accumulation of disability is related to demographic and disease-related variables, modeling of how the progression of EDSS varies from subject-to-subject, for example, remains unchanged, increases (disease worsening) or decreases (disease improvement), is of interest. In addition, similar to many other neuroinflammatory and neurodegenerative diseases, MS is heterogeneous, i.e., patients' EDSS scores change at individually varying rates, with a trend to gradually accumulate disability over the years.

A critical question for MS researchers is identifying the prognostic factors that influence EDSS progression. These factors provide insights into how progression varies between patients and point to possible directions for treatment. Different factors have been reported to be associated with the progression of disease; for instance, Vukusic and Confavreux (2007) showed age is a factor of progression; Runmarker and Andersen (1993) showed 5 years after onset of MS that a low number of affected neurological systems, a low neurological deficit score, and a high degree of remission from the last bout were the most important prognostic factors.

The analysis of disease progression in MS clinical trials has typically used time-to-event analysis, which is the recommended method for the demonstration of a drug effect on the first clinically meaningful disability progression in regulatory guidelines (Alvarez and others, 2015). Other works on exploring prognostic factors in MS (Kappos and others, 2015; Vukusic and Confavreux, 2003) have also relied on a time-to-event modeling framework. However, outside demonstrating a treatment effect on disease progression in randomized controlled trial settings, time-to-event models are not adequate for characterizing disease progression for three reasons; firstly, time-to-event models consider unidirectional changes and hence ignore that patients can improve or recover from their current EDSS state (see Mandel and others, 2007 for a detailed discussion), secondly, they do not account for repeated events of worsening thereby not fully using the available longitudinal information; and thirdly, these models cannot handle heterogeneity in EDSS transitions, i.e., time-to-event analysis treats worsening from EDSS 2 to 3 the same as worsening from EDSS 4 to 5. However, factors that influence the early transitions (e.g., from EDSS 2 to 3) may not be the same as the factors influencing transitions at a later stage in the disease (e.g., from EDSS 4 to 5). For this reason, survival models for EDSS are subject to biases introduced by left censoring, i.e., they do not account for the fact that patients in the analysis may start at different EDSS scores. Markov models address all of these concerns, allowing any transition (deterioration and recovery) and flexible modeling of covariates, where the covariate effect is estimated for each possible transition separately. A Markov process is one where the outcome at a given time only depends on the outcome of the previous time and is independent of its past history.

In this article, we focus on the continuous-time Markov model (CTMM). To aid us in defining this type of model, consider that Markov models can either be defined in discrete or continuous time. A discrete-time Markov model is one in which the system evolves through discrete time steps (e.g., every 1 month), while in a continuous-time Markov model transitions between states can happen at

any continuous time (e.g., 2 months, 1 year, 20 days etc.). MS patients often have measurements that are taken at variable time-points because in addition to regular check-ups, EDSS assessments are also conducted when patients relapse. An MS relapse is an acute neurological deterioration caused by a demyelinating event in the central nervous system, or in other words, a sudden worsening in MS symptoms, from which patients may or may not fully recover. These varying observation times, therefore, require a continuous time model. A second consideration is whether a patient's observation time corresponds to the exact time when a change in status occurs. The exact time that a change in disability (EDSS) occurs for an MS patient may occur before a patient arrives in the clinic. Therefore, our data are interval censored. Because the exact time that a transition from one disability state to another is not directly observed, the CTMM as outlined in [Kalbfleisch and Lawless \(1985\)](#) is a suitable choice for our EDSS data. This form of likelihood takes into consideration all possible intermediate state transitions and timings between observation time-points. Once it's been fit to the data, the CTMM can generate the probability of moving from one EDSS state to another at any time, allowing one to estimate how factors are associated with faster or slower EDSS transition. CTMM are widely utilized to fit multistate longitudinal data. These models in particular are applied in the field of public health, where the states of a Markov chain may refer to worsening stages of a chronic disease, such as breast cancer ([Hsieh and others, 2002](#)). In longitudinal settings, individuals are followed at regular intervals, where the exact transition times between disease states are typically not observed. Because state transitions can happen at any time and data are collected in a nonequidistant longitudinal manner, the CTMM offer a more parsimonious approach over the discrete version of Markov models, and handle the underlying variable nature of the disease. Lastly, previous work has applied hidden Markov Models to multistate data in continuous time for progression ([Williams and others, 2020](#); [Luo and others, 2021](#)). Herein, we however choose to focus on the non-hidden Markov model. This is because, in MS patients, the current EDSS state is generally the most prominent factor in determining a patients future disability status rather than a latent unobserved state.

A recent study ([Lublin and others, 2022](#)) showed that relapse and age are important factors in the accumulation of disability in MS disease, although there is continuous interest in which clinical factors are associated with EDSS transitions. Therefore, our goal here is to estimate the effect of different covariates on different EDSS transitions by applying CTMM to a large longitudinal data set of MS patients. We are motivated by the large-scale NO.MS data set, which includes approximately 20,000 MS patients with longitudinal EDSS data followed for up to 15 years ([Dahlke and others, 2021](#)), and with a range of clinically relevant covariates. However, the large sample size and the nonequidistant time intervals between outcome measurements present a challenge for a CTMM because the likelihood and score function involve the evaluation of a matrix exponential for each observation, and as a result, model fitting and parameter estimation become burdensome, requiring numerical approximation methods ([Moler and Van Loan, 2003](#); [Kalbfleisch and Lawless, 1985](#)). Previously published articles focus on models in which the number of EDSS states has been restricted, with three or fewer state transitions allowed ([Mandel and others, 2013](#)). However, in this article, we relax this assumption and allow for more granular transitions between different EDSS states. A large database allows us to explore the influence of covariates in specific stages of the MS disease course. [Mandel and others \(2007\)](#) developed a method for analyzing MS disease states using fixed-effects transition models, which along with their first-order Markov assumption suffered from lack of fit, partially due to the heterogeneous nature of the disease. NO.MS contains a large number of observations, allowing for more granular transitions between EDSS states, which results in a large number of model parameters, and the analysis becomes complex and computationally intense. In this article, we propose a mini-batch stochastic gradient descent optimization technique combined with differentiation of the matrix exponential, based on

the approach presented in [Van Loan \(1978\)](#) and [Wilcox \(1967\)](#). This method has been previously applied to the context of hidden Markov models ([Marshall and Jones, 1995](#)), but to our knowledge has not been utilized for CTMM. We show that the mini-batch stochastic gradient used in this article can accommodate large scale data. While other studies have modeled the transitions between different EDSS states without investigating clinical prognostic factors ([Zurawski and others, 2019](#)), these factors or covariates affect the transition intensities between the different disease states. Indeed, the introduction of covariates should allow increased accuracy when predicting transition rates ([Cook and others, 2002](#)). We show that our model can handle different types of covariates including baseline as well as time-varying covariates.

Section 2 represents the details of the model and inference. Section 3 presents our approach to estimation and inference in CTMM. Section 4 first describes how we simulate data from the proposed CTMM and then describes how we assess the performance of our models. Finally, Section 5 presents an application and discussion of the proposed model on the NO.MS.2 data set.

## 2. MODEL

### 2.1. Continuous time Markov models

Consider a longitudinal study in which individuals can move among  $\mathcal{S} = \{1, 2, \dots, S\}$  states. Let  $M$  denote the number of subjects in the study and  $N_m$  denote the number of observation times for the  $m$ th subject. The time of subject  $m$ 's  $k$ th observation is denoted by  $t_{mk}$ , when we observe the subject's state as a random variable denoted by  $s_{mk} \in \mathcal{S}$ . States are assumed to follow a first-order continuous time, discrete state Markov process, that is,

$$\Pr(s_{mk} | s_{m(k-1)}, s_{m(k-2)}, \dots, s_{m1}) = \Pr(s_{mk} | s_{m(k-1)}),$$

and we denote the probability of subject  $m$ 's transition from state  $i$  to state  $j$  as

$$p_{ij}(t_{m(k-1)}, t_{mk}) = \Pr(s_{mk} = j | s_{m(k-1)} = i).$$

Figure 1 shows an illustration of state transitions and their corresponding probabilities. Individuals, after staying for some time in state  $i$ , move with some probability  $p_{ij}$  from state  $i$  to state  $j$ . The likelihood for each individual is the product of all such transitions across observation times. The Markov process is fully characterized by its transition intensities

$$q_{ij}(t) = \lim_{\Delta t \rightarrow 0} \frac{p_{ij}(t, t + \Delta t)}{\Delta t}, \quad i \neq j,$$

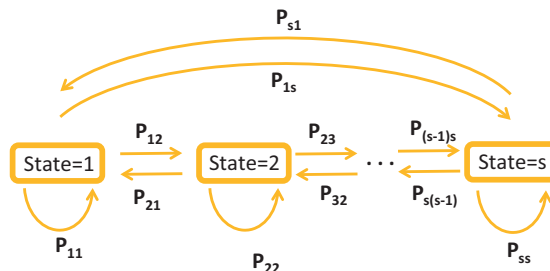


Fig. 1. Transitions between states and their corresponding probabilities

where  $q_{ii}(t) = -\sum_{j \neq i} q_{ij}(t)$  (Cox and Miller (1977)). For the matrix representation, suppose  $\mathbf{P}(t_1, t_2)$  and  $\mathbf{Q}(t)$  denote the  $S \times S$  matrix of transition probabilities  $p_{ij}(t_1, t_2)$  and matrix of transition intensities  $q_{ij}(t)$ , respectively (the transition intensity matrix is also known as the rate or infinitesimal generator matrix). Each entry  $q_{ij}(t)$  of the transition matrix  $\mathbf{Q}$  represents the rate of transition from state  $i$  to state  $j$  at time  $t$ .

Under time homogeneity, transition intensities are independent of the time, i.e.,  $q_{ij}(t) = q_{ij}$  and transition probabilities depend only on the elapsed time between successive observations, i.e.,  $p_{ij}(t_1, t_2) = p_{ij}(t_2 - t_1, 0)$  for  $i, j \in S$ . Then in this case, transition probabilities can be related directly to transition intensities through the so-called Kolmogorov equation,

$$\mathbf{P}(t_1, t_2) = \text{Exp}(\mathbf{Q}(t_2 - t_1)), \quad (2.1)$$

where  $\text{Exp}(\cdot)$  indicates the matrix exponential (we denote the scalar exponential with  $\exp(\cdot)$ ).

**2.1.1. Incorporating covariates in the model** Suppose subject  $m$  has  $R$  covariates  $z_{mr}(t_{mk})$ ,  $r = 1, \dots, R$ , measured at time  $t_{mk}$  ( $k$ th observation), and written as the  $R$ -vector  $\mathbf{Z}_m(t_{mk})$ . This could contain different types of covariates such as time-independent covariates (e.g., sex), as well as time-varying ones (e.g., age). We further assume that the covariate value remains constant in the future from the present value until the next observation. Marshall and Jones (1995) described a form of proportional hazards model, wherein the presence of covariates, the transition intensity matrix elements  $q_{ij}$  are replaced by

$$q_{ij,m}(\mathbf{Z}_m(t_{mk})) = q_{ij}^0 \exp\left(\sum_{r=1}^R z_{mr}(t_{mk}) \beta_{ij,r}\right), \quad (2.2)$$

where  $\beta_{ij,r}$  is a regression coefficient that quantifies the impact of each covariate on state transition from  $i$  to  $j$ , and  $q_{ij}^0$  is the baseline intensity; we write  $\beta$  for the  $R \times S \times (S - 1)$ -vector of all coefficients, and  $\mathbf{Q}^0$  is called baseline intensity matrix with entries  $q_{ij}^0$ . The interpretability of  $\mathbf{Q}^0$  is the transition intensity when all covariates take on the value zero or at the reference value for categorical covariates, and as a result, all covariates should generally be centred. We call this version of our model a “transition-dependent regression” where the effect of each covariate can vary for each state transition. We also can consider another version where regressors have a common effect over all state transitions — i.e.,  $q_{ij,m}(\mathbf{Z}_m(t_{mk})) = q_{ij}^0 \exp\left(\sum_{r=1}^R z_{mr}(t_{mk}) \beta_r\right)$ , which we call the transition-independent regression case. For the rest of the article, we will focus on the transition-dependent regression, though the transition-independent regression is modeled similarly and is nested within the more flexible model.

Let  $\tau_{mk} = t_{mk} - t_{m(k-1)}$ ,  $k = 1, \dots, N_m$ , be the time lag between two consecutive observation times. Denote the parameters of interest  $\theta$ , elements of  $\mathbf{Q}^0$  and  $\beta$ , with dimension  $R \times S \times (S - 1)$ . This allows us to formulate the time-invariant likelihood for subject  $m$  as follows

$$L_m(\theta) = \prod_{k=1}^{N_m} \left( \mathbf{P}_{mk}(t_{mk}, t_{m(k-1)}) \right)_{s_{m(k-1)} s_{mk}} = \prod_{k=1}^{N_m} \left( \text{Exp}(\mathbf{Q}_{mk} \tau_{mk}) \right)_{s_{m(k-1)} s_{mk}}, \quad (2.3)$$

where  $(\cdot)_{s_{m(k-1)} s_{mk}}$  refers to the corresponding entry of the matrix (row: state at time  $t_{m(k-1)}$ , and column: state at time  $t_{mk}$ ). The complete likelihood then becomes the product over all subjects, i.e.,  $L(\theta) = \prod_{m=1}^M L_m(\theta)$ . Our likelihood in (2.3) is the generalized version of likelihood used

by Kalbfleisch and Lawless (1985) when covariate vectors are measured at every observation time (canonical decomposition of  $\mathbf{Q}$  at every observation).

The ML estimator of  $\theta$  can be obtained through the maximization of  $L(\theta)$ . Although  $L(\theta)$  has a simple form, evaluation is computationally intensive because it includes the matrix exponential operation for each product of the likelihood. In the next section, we introduce our approach to calculating the derivatives of  $L(\theta)$  and develop a stochastic gradient descent approach to find the ML estimates.

### 3. ESTIMATION AND INFERENCE

#### 3.1. Optimization with stochastic gradient descent

Combined together, Wilcox (1967) and Van Loan (1978) (Theorem 1) show that for any matrix function  $\mathbf{A}(\lambda) = (a_{ij}(\lambda))$  for scalar  $\lambda$ , the derivative of  $\text{Exp}(\mathbf{A}(\lambda))$  w.r.t  $\lambda$  can be found via a Padé approximation

$$\frac{\partial}{\partial \lambda} \text{Exp}(\mathbf{A}(\lambda)) = \int_0^1 \text{Exp}(u\mathbf{A}) \dot{\mathbf{A}} \text{Exp}((1-u)\mathbf{A}) du \approx \left( \text{Exp}(\mathbf{C}) \right)_{1:S, (S+1):(2S)}, \quad (3.4)$$

where  $\dot{\mathbf{A}} = (\dot{a}_{ij}(\lambda))$ ,  $\dot{a}_{ij}(\lambda) = \partial a_{ij}(\lambda) / \partial \lambda$  (i.e., element-wise derivative),  $\mathbf{C} = \begin{pmatrix} \mathbf{A} & \dot{\mathbf{A}} \\ \mathbf{0} & \mathbf{A} \end{pmatrix}$ ,  $\mathbf{0}$  is  $S \times S$  zero matrix, and subscripts  $(\cdot)_{1:S, (S+1):(2S)}$  indicate extraction of the upper right  $S \times S$  submatrix.

Our goal is to obtain the partial derivatives of  $L(\theta)$  w.r.t elements of  $\theta$ , with  $\theta_\ell$  standing in for each  $q_{ij}^0$  and  $\beta_{ij,r}$ , for all transitions  $i, j \in \mathcal{S}$  such that  $i \neq j$  and all covariates  $r$ ; for subject  $m$

$$\frac{\partial}{\partial \theta_\ell} \log(L_m(\theta_\ell)) = \sum_{k=1}^{N_m} \frac{\partial}{\partial \theta_\ell} \log(\text{Exp}(\mathbf{Q}_{mk} \tau_{mk})) = \sum_{k=1}^{N_m} \frac{\partial \text{Exp}(\mathbf{Q}_{mk} \tau_{mk}) / \partial \theta_\ell}{\text{Exp}(\mathbf{Q}_{mk} \tau_{mk})}, \quad (3.5)$$

where the final ratio of matrices is evaluated as an entry-wise ratio for each entry  $(i, j)$ ,  $i \neq j$ . Thus, we need to calculate the derivatives for every subject  $m$  and at every observation time lag  $\tau_{mk}$ . To use the result (3.4) above, for each event  $\tau_{mk}$ , we take  $\mathbf{A} = \mathbf{Q}_{mk} \tau_{mk}$  which gives  $\dot{\mathbf{A}} = \dot{\mathbf{Q}}_{mk} \tau_{mk}$ , the partial derivatives w.r.t  $q_{ij}^0$ ,

$$\frac{\partial}{\partial q_{ij}^0} (\mathbf{Q}_{mk} \tau_{mk}) = \exp(\beta_{ij,1} z_{m1}(\tau_{mk}) + \beta_{ij,2} z_{m2}(\tau_{mk}) + \cdots + \beta_{ij,r} z_{mr}(\tau_{mk})) \tau_{mk},$$

and elements corresponding to derivatives w.r.t  $\beta_{ij,r}$ ,

$$\frac{\partial}{\partial \beta_{ij,r}} (\mathbf{Q}_{mk} \tau_{mk}) = q_{ij}^0 z_{mr}(\tau_{mk}) \times \exp(\beta_{ij,1} z_{m1}(\tau_{mk}) + \cdots + \beta_{ij,r} z_{mr}(\tau_{mk})) \tau_{mk}.$$

According to (3.4), having the matrices  $\dot{\mathbf{A}}$  for each observation time lag, we can form the block matrices  $\mathbf{C}$  and use (3.5) to calculate  $\partial / \partial \theta \log(L_m(\theta))$ .

Finally, the gradient is obtained,

$$\frac{\partial}{\partial \theta} \log(L(\theta)) = \sum_{m=1}^M \frac{\partial}{\partial \theta} \log(L_m(\theta)). \quad (3.6)$$

Having  $\partial / \partial \theta \log(L(\theta))$ , we can now use gradient descent to find ML estimates of the parameters  $\theta$  in (2.3). However, the standard gradient descent requires application on the full data set, hence,



computing matrix exponentials and derivatives at each observation time lag  $\tau_{mk}$  for all individuals. Consequently, it is computationally expensive to scale it to a large data set. Instead, we use a mini-batch stochastic version of gradient descent (Sakrison, 1965), which randomly partitions the data into mini-batches and performs the calculations on each mini-batch instead of the entire data set. This allows our method to scale to essentially arbitrarily large data sets. Let  $\mathcal{B}_d$  be a randomly chosen subset of subjects (with or without replacement) at the iteration step  $d$  ( $|\mathcal{B}_d| < M$ ). The updated parameters at step  $d + 1$  of the mini-batch stochastic gradient descent are obtained via

$$\theta_{d+1} = \theta_d + \lambda_{d+1} \frac{M}{|\mathcal{B}_d|} \sum_{m \in \mathcal{B}_d} \left( \frac{d}{d\theta} \log L_m(\theta) \right) \Big|_{\theta=\theta_d}, \quad (3.7)$$

where  $\lambda_{d+1}$  is the learning rate sequence (we use decreasing learning rate with starting value equal to  $\lambda_{d+1} = (d+1)^{-0.6}$ ). This procedure iterates until the parameters using (3.7) converge to  $\hat{\theta}$  (i.e.,  $\hat{\theta}_{d+1} - \hat{\theta}_d < \epsilon$  for a small  $\epsilon$ ). This then results in ML estimators of  $\hat{\theta} = (\{q^0\}_{ij}, \{\beta\}_{ij,r})$ , for all transitions  $i, j \in \mathcal{S}$ ,  $i \neq j$ , and all covariates  $r$ . Moving forward, we will refer to this optimization approach as the SCTMM.

**3.1.1. Calculation of confidence intervals** Asymptotic standard errors and confidence intervals are computed from the Hessian matrix of the log-likelihood evaluated at the parameter estimates. There are different numerical approximations/approaches in the literature to compute the Hessian for the parameters  $\theta$ . Since the calculation of the Hessian matrix requires computing  $(S \times (S-1) \times R)^2$  elements, existing methods (Hanks, 2018; Fleming and Calabrese, 2021; Jackson, 2011) are computationally prohibitive if  $M$  is large. Here, we propose two scalable approaches to calculate the Hessian matrix; the first is via calculation of second-order derivatives through reapplication of the Padé approximation in (3.4); and the second approach is with a second-order approximation using the power series definition of the matrix exponential (see Moler and Van Loan, 2003).

### Padé expansion for Hessian

Let  $\theta = (\theta_1, \theta_2)$  where  $\theta_1$  and  $\theta_2$  are the  $\{q^0\}_{ij}$  and  $\{\beta\}_{ij,r}$  components. The Hessian then has a  $2 \times 2$  block form, with blocks  $(\ell, \ell')$ , with  $\ell, \ell' \in \{1, 2\}$ . To obtain the Hessian matrix, we need to calculate the second derivative of the log-likelihood stated in (2.3) as follows

$$\begin{aligned} \frac{\partial^2}{\partial \theta_\ell \partial \theta_{\ell'}} \log L(\theta) &= \frac{\partial^2}{\partial \theta_\ell \partial \theta_{\ell'}} \log \left( \prod_{m=1}^M \prod_{k=1}^{N_m} (\text{Exp}(\mathbf{Q}_{mk} \tau_{mk}))_{s_{m(k-1)}, s_{mk}} \right) \\ &= \sum_{m=1}^M \sum_{k=1}^{N_m} \frac{\partial}{\partial \theta_{\ell'}} \left( \frac{\left( \frac{\partial}{\partial \theta_\ell} \text{Exp}(\mathbf{Q}_{mk} \tau_{mk}) \right)_{s_{m(k-1)}, s_{mk}}}{(\text{Exp}(\mathbf{Q}_{mk} \tau_{mk}))_{s_{m(k-1)}, s_{mk}}} \right) \\ &= \sum_{m=1}^M \sum_{k=1}^{N_m} \frac{\left( \frac{\partial}{\partial \theta_{\ell'}} \left( \frac{\partial}{\partial \theta_\ell} \text{Exp}(\mathbf{Q}_{mk} \tau_{mk}) \right) \right)_{s_{m(k-1)}, s_{mk}} (\text{Exp}(\mathbf{Q}_{mk} \tau_{mk}))_{s_{m(k-1)}, s_{mk}}}{(\text{Exp}(\mathbf{Q}_{mk} \tau_{mk}))_{s_{m(k-1)}, s_{mk}}^2} \\ &\quad - \frac{\left( \frac{\partial}{\partial \theta_{\ell'}} \text{Exp}(\mathbf{Q}_{mk} \tau_{mk}) \right)_{s_{m(k-1)}, s_{mk}} \left( \frac{\partial}{\partial \theta_\ell} \text{Exp}(\mathbf{Q}_{mk} \tau_{mk}) \right)_{s_{m(k-1)}, s_{mk}}}{(\text{Exp}(\mathbf{Q}_{mk} \tau_{mk}))_{s_{m(k-1)}, s_{mk}}^2}. \end{aligned} \quad (3.8)$$

The terms in expression (3.8) can be found in Section 3.1, except for  $\partial/\partial\theta_{\ell'}(\partial/\partial\theta_{\ell}\text{Exp}(\mathbf{Q}_{mk}\tau_{mk}))$ , which is found as follows

$$\begin{aligned} \frac{\partial}{\partial\theta_{\ell'}}\left(\frac{\partial}{\partial\theta_{\ell}}\text{Exp}(\mathbf{Q}_{mk}\tau_{mk})\right) &= \frac{\partial}{\partial\theta_{\ell'}}\left(\left(\text{Exp}\left(\begin{smallmatrix} \mathbf{Q}_{mk}\tau_{mk} & \frac{\partial}{\partial\theta_{\ell}}(\mathbf{Q}_{mk}\tau_{mk}) \\ \mathbf{0} & \mathbf{Q}_{mk}\tau_{mk} \end{smallmatrix}\right)\right)_{1:S,(S+1):(2S)}\right) \\ &= \left(\frac{\partial}{\partial\theta_{\ell'}}\text{Exp}\left(\begin{smallmatrix} \mathbf{Q}_{mk}\tau_{mk} & \frac{\partial}{\partial\theta_{\ell}}(\mathbf{Q}_{mk}\tau_{mk}) \\ \mathbf{0} & \mathbf{Q}_{mk}\tau_{mk} \end{smallmatrix}\right)\right)_{1:S,(S+1):(2S)} \\ &= \left(\left(\text{Exp}\left(\begin{smallmatrix} \mathbf{C}_1 & \mathbf{C}_2 \\ \mathbf{C}_3 & \mathbf{C}_1 \end{smallmatrix}\right)\right)_{1:2S,(2S+1):(4S)}\right)_{1:S,(S+1):(2S)} \\ &= \left(\text{Exp}\left(\begin{smallmatrix} \mathbf{C}_1 & \mathbf{C}_2 \\ \mathbf{C}_3 & \mathbf{C}_1 \end{smallmatrix}\right)\right)_{1:S,(3S+1):(4S)}, \end{aligned} \quad (3.9)$$

where  $\partial/\partial\theta_{\ell}(\mathbf{Q}_{mk}\tau_{mk}) = \partial\{q\}_{ij}\tau_{mk}/\partial\theta_{\ell}$ ,  $\partial^2/\partial\theta_{\ell'}\partial\theta_{\ell}(\mathbf{Q}_{mk}\tau_{mk}) = \partial^2\{q\}_{ij}\tau_{mk}/\partial\theta_{\ell'}\partial\theta_{\ell}$ ,  $\mathbf{C}_1 = \begin{pmatrix} \mathbf{Q}_{mk}\tau_{mk} & \frac{\partial}{\partial\theta_{\ell}}(\mathbf{Q}_{mk}\tau_{mk}) \\ \mathbf{0} & \mathbf{Q}_{mk}\tau_{mk} \end{pmatrix}$ ,  $\mathbf{C}_2 = \begin{pmatrix} \frac{\partial}{\partial\theta_{\ell'}}(\mathbf{Q}_{mk}\tau_{mk}) & \frac{\partial^2}{\partial\theta_{\ell'}\partial\theta_{\ell}}(\mathbf{Q}_{mk}\tau_{mk}) \\ \mathbf{0} & \frac{\partial}{\partial\theta_{\ell'}}(\mathbf{Q}_{mk}\tau_{mk}) \end{pmatrix}$ ,  $\mathbf{C}_3$  is  $2S \times 2S$  zero matrix, and  $\mathbf{0}$  is an  $S \times S$  zero matrix. Specification of the Hessian is completed by considering the different possible values for  $\theta_{\ell'}$  and  $\theta_{\ell}$ :

- (1)  $\theta_{\ell'} = q_{ij}^0, \theta_{\ell} = q_{ij}^0 : \frac{\partial^2}{\partial\theta_{\ell'}\partial\theta_{\ell}}\mathbf{Q}_{mk}\tau_{mk} = 0$ ,
- (2)  $\theta_{\ell'} = q_{ij}^0, \theta_{\ell} = \beta_{ij,r}$  or  $\theta_{\ell'} = \beta_{ij,r}, \theta_{\ell} = q_{ij}^0 : \frac{\partial^2}{\partial\theta_{\ell'}\partial\theta_{\ell}}\mathbf{Q}_{mk}\tau_{mk} = z_{mr}(\tau_{mk}) \sum_{r=1}^R z_{mr}(\tau_{mk})\beta_{ij,r}$ ,
- (3)  $\theta_{\ell'} = \beta_{ij,u}, \theta_{\ell} = \beta_{ij,r} : \frac{\partial^2}{\partial\theta_{\ell'}\partial\theta_{\ell}}\mathbf{Q}_{mk}\tau_{mk} = q_{ij}^0 z_{mr}(\tau_{mk})z_{mu}(\tau_{mk}) \sum_{r=1}^R z_{mr}(\tau_{mk})\beta_{ij,r}$ , for every  $u$ .

### Power series expansion for Hessian

Using the Padé approximation to calculate confidence intervals is computationally expensive especially when number of states  $S$  is large, due to the costly computation of the Exp function. Thus, we also introduce a second approach using approximation, which could be faster in terms of computation but with a slight compromise in accuracy. Recalling (2.3) and using the definition of matrix exponential we have

$$\begin{aligned} \log L_m(\theta) &= \sum_{k=1}^{N_m} \log(\text{Exp}(\mathbf{Q}_{mk}\tau_{mk}))_{s_m(k-1)s_{mk}} \\ &= \sum_{k=1}^{N_m} \log\left(\sum_{l=0}^{\infty} \frac{1}{l!}(\mathbf{Q}_{mk}\tau_{mk})^l\right)_{s_m(k-1)s_{mk}} \\ &\approx \sum_{k=1}^{N_m} \log(\mathbf{I} + \mathbf{Q}_{mk}\tau_{mk} + 1/2(\mathbf{Q}_{mk}\tau_{mk})^2)_{s_m(k-1)s_{mk}}, \end{aligned} \quad (3.10)$$

where  $\mathbf{I}$  is the identity matrix,  $s_{mk}$  is the occupied state at time  $t_{mk}$ , and we truncate the power series after the third term (see Appendix for more details). We need to mention that the trade-off in this method is that truncation leads to a slight underestimation of the variance in the estimation of confidence intervals since some of the terms in the power series are dropped. We will explore the



impact on coverage in Section 4. Having the second derivatives we can simply form the Hessian matrix and the calculation of confidence intervals then becomes straightforward.

### 3.2. Evaluation methods

We compare our method to an existing and widely used CTMM tool. [Jackson \(2011\)](#) proposed a method that allows CTMM to be fitted to longitudinal data and developed its R package called MSM as well ([Jackson, 2011](#)). MSM uses different approaches for the optimization step, and we will evaluate each individually:

- MSM\_opt: optim method which uses the deterministic Nelder-mead approach ([R Core Team, 2021](#)).
- MSM\_nlm: Which uses Newton-type algorithm ([R Core Team, 2021](#)).
- MSM\_F: Which uses fisher scoring ([Kalbfleisch and Lawless, 1985](#)).

## 4. SIMULATION STUDY

### 4.1. Simulation overview

We perform a simulation study under different scenarios to assess the performance of both the proposed methods for optimization by SCTMM and the construction of confidence intervals outlined in Section 3. Let  $t_{\max}$  be the maximum time in which subjects are followed up (for instance 10 years). Simulating data for the CTMM for one subject ( $m$ ) is carried out via the following steps:

1. We first choose an arbitrary baseline transition matrix  $\mathbf{Q}^0$  and some values for the baseline covariates  $z_{mr}(t_{m0})$ , for every covariate  $r$ , then we can compute  $\mathbf{Q}_{m0}$  via equation (2.2). The interpretation of the  $\mathbf{Q}^0$  depends on the centering of the covariates. To obtain the stationary (also called steady) state probability (shown by  $\pi$ ), we solve the following matrix equation:

$$\pi^T \mathbf{Q} = 0.$$

This specifies the initial state occupied by subject  $m$ . In other words, we draw a sample state from the finite space of states  $\mathcal{S}$  with probabilities  $\pi$ .

2. To obtain the time of the next state transition  $t_{m(k+1)}$ , we draw one sample from an exponential distribution with rate  $-q_{s_{mk}s_{mk}}$ , where  $s_{mk}$  stands for the state occupied by subject  $m$  at observation  $k$  (note that for the very first step of the simulation algorithm  $k = 0$ ). Furthermore, randomly generate some values from a uniform distribution for  $z_{mr}(t_{m(k+1)})$  for every  $r$ .
3. Notice that the time  $t_{m(k+1)}$  generated at Step 2, is the instantaneous transition time where subject  $m$  moves to any other state but not the currently occupied one  $s_{mk}$ . Because of this, in order to simulate patients who remain clinically stable, we need to generate a number of dummy observations in which subject  $m$  has been observed several times staying at the same state  $s_{mk}$  but with different values of covariates (randomly). The choice of how many dummy sojourn observations would be required is arbitrary; however, it is advised to use the frequency in which the states are observed in the real data set at hand to mimic the data; for instance, every 4 months.

4. Now, we choose the next state  $s_{m(k+1)}$  by drawing a sample from the set of states  $\mathcal{S}$  with probabilities computed by equation (2.1).
5. If  $t_{m(k+1)} < t_{\max}$  go to step 2, otherwise terminate the algorithm.

We then repeat the above steps for any given number of subjects  $M$ .

#### 4.2. Simulation

We consider two main simulation scenarios; first, a null case where there is no effect of covariates on the transition rates (i.e.,  $\beta_{ij,r} = 0$  for every  $i, j$ , and  $r$ ); and a second case where there is an effect from the covariate impacting transition rates. In both cases, we try to mimic the real data set that we will use in the next section and choose the number of states based on where we have the majority of transition data, and so we reduce the 20 EDSS states down to eight states ( $\mathcal{S} = \{1, 2, \dots, 8\}$ ),  $t_{\max} = 15$  years (follow-up time). For the second case where the covariates impact the transition rates, we use 10 covariates and assess the estimation of transition-dependent regression parameters  $\beta_{ij,r}$ . All 10 covariates are randomly drawn from a uniform distribution and then centered at zero. We perform a sensitivity analysis over sample size  $M$  in both scenarios and assess the performance of our SCTMM method. For both scenarios, and each sample size, we perform Monte Carlo simulation and generate 1000 realizations of the data set, and evaluate the operating characteristics of bias, standard error, coverage, and rejection rates. Estimation with each realization uses different initial values for  $\beta_{ij,r}$  for each  $i, j$ , and  $r$ . For each realization,  $\beta_{ij,r}$  are drawn from a normal distribution with mean 0 and standard deviation 1. The bench-marking platform used for this study ran R-4.0.0 to generate data sets and perform the analysis on 7 Intel Ivy Bridge cores each running at 1.15 GHz speed and 16 GB of RAM memory in total.

There are many configuration choices in implementing a mini-batch stochastic gradient optimization, and we take guidance from the existing literature (Franchini and others, 2020; Perrone and others, 2019). With large-scale data ( $M > 1000$ ), it is advised to set the mini-batch size  $|\mathcal{B}_d|$  to be between 500 and 1000. In this article, we fix  $|\mathcal{B}_d| = 500$ .

The CTMM likelihood is not globally concave and thus multiple restarts are recommended to get as close to the global maximum as possible; the optimization result that produces the largest log likelihood is taken as the optimal solution. Thus, we advise using between 1000 and 5000 restarts with random initial values, for any number of parameters up to what we have here (we use 1000 restarts). Careful selection of initial values is required to reduce the chance of the optimizer arriving at saddle points and local optimal. We suggest one of the following policies where applicable to set the starting parameters for the SCTMM method:

- Apply SCTMM on a smaller section of the data with no stochastic computation (for instance  $M = 500$ ), estimate the parameters and then set these derived parameters for the restarts in the SCTMM.
- Sample off-diagonal elements for row  $i$  and column  $j$  from positive values of a univariate normal distribution with mean and standard deviation  $1/|i - j|$ .

The above suggestions aim at minimizing the number of restarts leading to highly suboptimal solutions and improving the accuracy of the estimates. We perform two analyses, one with no covariates in the model, where we estimate the transition intensities across all state transitions, and a second case where we include covariates and remove from the model any state transition  $ij$  that has less than 1% transition data in the entire data set. Then, we estimate covariate effects for this case. We also use the Padé expansion method to calculate confidence intervals.

### 4.3. Simulation Results

To assess the estimation performance of the coefficients  $\beta_{ij,r}$  in the presence of covariate effects (i.e., the transition-dependent scenario), we have tested the performance of SCTMM against MSM. Table 1 shows a comparison between the optimization methods used in MSM to obtain maximum likelihood estimates, and our proposed SCTMM method. These results show that under the small-scale setting ( $M \leq 1000$ ) SCTMM has equal or better performance to MSM in terms of bias, variance, coverage, and rejection rate. While all methods have some under coverage and slight inflation of the null rejection rate, SCTMM is closest to the nominal. In a large-scale setting ( $M \geq 10000$ ), none of MSM's methods was available (due to numerical instability), but the SCTMM approach estimates the parameters with reasonably good performance. Furthermore, the results demonstrate the scalability of the SCTMM approach to large data sets. Table 2 shows a comparison between the optimization methods used in the MSM software and our proposed SCTMM approach when there is no effect of any covariates incorporated in the model ( $\beta_{ij,r} = 0$  for every  $i, j$ , and  $r$ ). Again, the results show that under a small-scale setting, SCTMM has equal or better performance than MSM, and in a large-scale setting the SCTMM can handle the scalability problem. Figure 2 shows a comparison of the computation time between the SCTMM and MSM, for only one initialization (no restarts), demonstrating that both methods are on par in this case.

## 5. APPLICATION ON NO.MS.2 DATA SET

### 5.1. NO.MS.2 data

We utilize the novel Novartis-Oxford MS (NO.MS.2) data set in this article, formed by aggregating multiple sclerosis patient data from 34 MS clinical trials and their extensions (Dahlke and others, 2021; Lublin and others, 2022). Currently, NO.MS.2 is the largest and most comprehensive clinical trial data set in MS, spanning all MS phenotypes and containing data from over 27 000 patients with up to 15 years of follow-up visits, and with regular monitoring of patients' neurological status by highly trained raters across all stages of MS disease. The earliest clinical trial in this data set began in 2003 and the latest measurement is from January 2020. These trials collected longitudinal data on disability, as measured by Kurtzke's EDSS (Kurtzke, 1983). In this analysis, EDSS was simplified as follows: state 0 remained 0; for EDSS 1–1.5, 2–2.5, 3–3.5,  $\dots$ , were rounded down to 1, 2, 3,  $\dots$ ;  $\text{EDSS} \geq 8$  was set to 8, resulting in 9 reduced states. In this analysis, EDSS was simplified as follows: state 0 remained 0; for EDSS 1–1.5, 2–2.5, 3–3.5,  $\dots$ , were rounded down to 1, 2, 3,  $\dots$ ;  $\text{EDSS} \geq 8$  was set to 8, resulting in 9 reduced states.

In this study, we used a cohort of  $M = 13\,320$  patients, with  $N = 170\,628$  observations in total,  $S = \{0, 1, \dots, 8\}$  EDSS states and a set of the following  $R = 7$  covariates.

- **ARR1:** The annualized relapse rate is the number of relapses a patient experiences within the year prior to the current observation time (time-varying).
- **Age:** Age at event time divided by 10, to interpret coefficients as the effect by a decade of age and to make coefficients more comparable with other variables (time-varying).
- **DURFS:** Duration of MS in years from first symptoms as estimated at trial entry (baseline).
- **MSTYPE:** A three-level categorical variable, indicating the phenotype of MS disease, with levels relapsing–remitting MS RRMS (used as reference level), secondary progressive MS (SPMS), and primary progressive MS (PPMS).
- **Sex:** Male or female (reference level female).

Table 1. Comparison of bias, variance, coverage, and rejection rate between the methods used in the MSM software and our proposed SCTMM method, when varying  $\beta_{ij,r}$ , sample size  $M$  and number of observations  $N$ .

M	N	True $\beta_{ij,r}$	Transition-dependent												Rejection rate											
			Bias				Variance				Coverage															
			MSM_opt	MSM_nlm	MSM_F	SCTMM	MSM_opt	MSM_nlm	MSM_F	SCTMM	MSM_opt	MSM_nlm	MSM_F	SCTMM	MSM_opt	MSM_nlm	MSM_F	SCTMM	MSM_opt	MSM_nlm	MSM_F	SCTMM	MSM_opt	MSM_nlm	MSM_F	SCTMM
500	15663	0	0.01	0.12	0.011	0.009	0.006	0.008	0.006	0.004	0.928	0.9	0.920	0.932	0.072	0.075	0.073	0.068								
500	15663	0.5	0.012	0.14	0.12	0.01	0.009	0.014	0.01	0.009	0.93	0.9	0.93	0.94	0.71	0.72	0.70	0.70								
500	15663	2	0.022	0.23	0.022	0.015	0.009	0.02	0.01	0.01	0.92	0.91	0.89	0.93	0.81	0.8	0.81	0.83								
1000	29891	0	0.01	0.12	0.11	0.009	1e-5	0.013	1e-4	1e-5	0.929	0.91	0.921	0.936	0.071	0.08	0.073	0.064								
1000	29891	0.5	0.005	0.10	0.009	0.004	0.009	0.016	0.009	0.009	0.93	0.89	0.9	0.94	0.71	0.68	0.69	0.70								
1000	29891	2	0.01	0.10	0.10	0.008	0.009	0.016	0.009	0.01	0.92	0.88	0.91	0.92	0.80	0.79	0.80	0.82								
10000	312547	0				0.008				0.005				0.91			0.09									
10000	312547	0.5				0.010				0.009				0.94			0.72									
10000	312547	2				0.010				0.009				0.93			0.83									
20000	619720	0				0.010				0.02				0.92			0.08									
20000	619720	0.5				0.010				0.02				0.91			0.73									
20000	619720	2				0.017				0.013				0.90			0.81									

Table 2. Comparison of bias, variance, coverage, and rejection rate, in the setting where there are no covariates, between the methods used in the MSM software and our proposed SCTMM method, when varying  $q_{ij}^0$ , sample size  $M$  and number of observations  $N$ .

Estimation of baseline intensities in absence of covariates ( $\beta_{ij,r} = 0$ )																										
			Bias					Variance					Coverage					Rejection rate								
$M$	$N$	True $q_{ij}^0$	MSM_opt	MSM_nlm	MSM_F	SCTMM	MSM_opt	MSM_nlm	MSM_F	SCTMM	MSM_opt	MSM_nlm	MSM_F	SCTMM	MSM_opt	MSM_nlm	MSM_F	SCTMM	MSM_opt	MSM_nlm	MSM_F	SCTMM	MSM_opt	MSM_nlm	MSM_F	SCTMM
500	15663	0	7e-6	8e-6	7e-6	7e-6	5e-6	7e-6	5e-6	6e-6	0.952	0.949	0.952	0.953	0.01	0.05	0.01	0.01	0.01	0.05	0.01	0.01	0.01	0.01	0.01	0.01
500	15663	0.5	9e-6	2e-5	9e-6	9e-6	7e-6	9e-6	7e-6	7e-6	0.951	0.949	0.952	0.951	0.73	0.71	0.72	0.73	0.73	0.71	0.72	0.72	0.73	0.72	0.73	
500	15663	1	1e-5	3e-5	4e-5	1e-5	7e-6	1e-6	7e-6	7e-6	0.94	0.939	0.941	0.94	0.84	0.8	0.83	0.89	0.89	0.8	0.83	0.83	0.89	0.83	0.89	
1000	29891	0	2e-5	4e-5	2e-5	1e-5	9e-6	2e-5	9e-6	8e-6	0.947	0.94	0.948	0.949	0.08	0.1	0.09	0.01	0.01	0.08	0.1	0.09	0.01	0.09	0.01	
1000	29891	0.5	4e-5	6e-5	4e-5	3e-5	3e-5	8e-5	2e-5	2e-5	0.949	0.94	0.952	0.946	0.77	0.75	0.77	0.79	0.79	0.75	0.77	0.77	0.79	0.77	0.79	
1000	29891	1	5e-5	7e-5	4e-5	4e-5	8e-5	1e-4	8e-5	7e-5	0.936	0.931	0.935	0.941	0.81	0.73	0.82	0.87	0.87	0.73	0.82	0.82	0.87	0.82	0.87	
10000	312547	0				0.001				5e-4			0.94					0.012							0.012	
10000	312547	0.5				0.003				0.003			0.937					0.78							0.78	
10000	312547	1				0.008				0.006			0.931					0.89							0.89	
20000	615720	0				0.009				0.007			0.941					0.02							0.02	
20000	615720	0.5				0.01				0.009			0.935					0.90							0.90	
20000	615720	1				0.02				0.01			0.93					0.88							0.88	

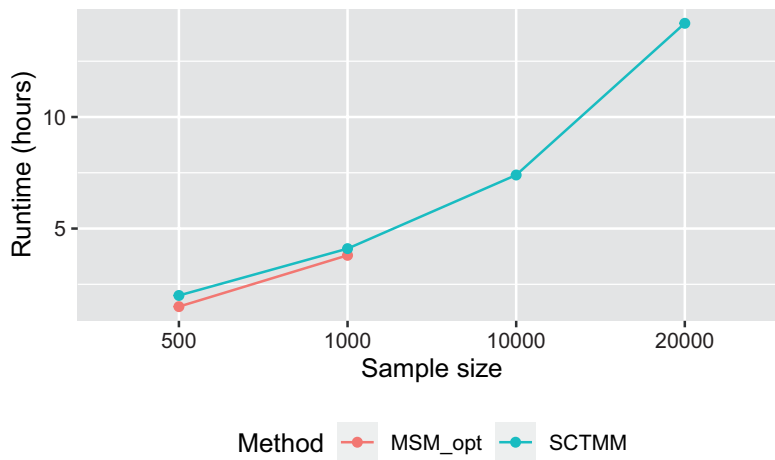


Fig. 2. Running time comparison between MSM software and the proposed SCTMM method.

We perform two analyses:

1. First, an analysis where we include no covariates in the model, and we consider all EDSS transitions (i.e., keeping rare transitions).
2. Second, an analysis where we only allow transitions that have more than 1% of transition data in the entire data set, which for this particular data set turns out to be mostly EDSS transitions with jumps of more than two states. For these transitions, we fit covariate effects.

All covariates are centered, and we apply SCTMM using mini-batches of size  $|\mathcal{B}_d| = 500$  in each iteration step  $d$ . We also use 1000 restarts with different initial values of  $\theta$ , and draw these initial values from a uniform distribution with  $\min = 0, \max = 1$  and  $\min = -1, \max = 1$  for  $\{q^0\}_{ij}$  and  $\{\beta\}_{ij,r}$ , respectively.

### 5.2. Real Data Results

The right panel of Figure 3 shows the model-based estimates of the baseline transition probability matrix when no covariate data were included in the model parameterization. It can be seen that patients are most likely to stay unchanged in their current states and then the probability of transitioning to a higher/lower EDSS state is higher in the early-mid stages of disability than in later stages of the disease. The plausibility of these results has been checked by looking at the left panel of Figure 3, which displays a normalized empirical transition probability matrix. This empirical transition matrix is calculated using adjacent clinical assessments in the entire dataset.

The graphs shown in Figure 4 illustrate transition-specific hazard ratios ( $\exp(z_r \beta_{ij,r})$ ), when the covariate is set to its mean value, and corresponding 95% confidence intervals (all covariates are included in the model at the same time). These graphs show the effect of each covariate on successive EDSS transitions. Figure 4a demonstrates hazard ratios of ARR1, showing that relapses have a significant association with the accumulation of EDSS disability, particularly in but not limited to, the early stages of the disease. This is in concordance with the findings of a recent study by Lublin and others (2022). Based on these results, for instance, having a relapse in the past year



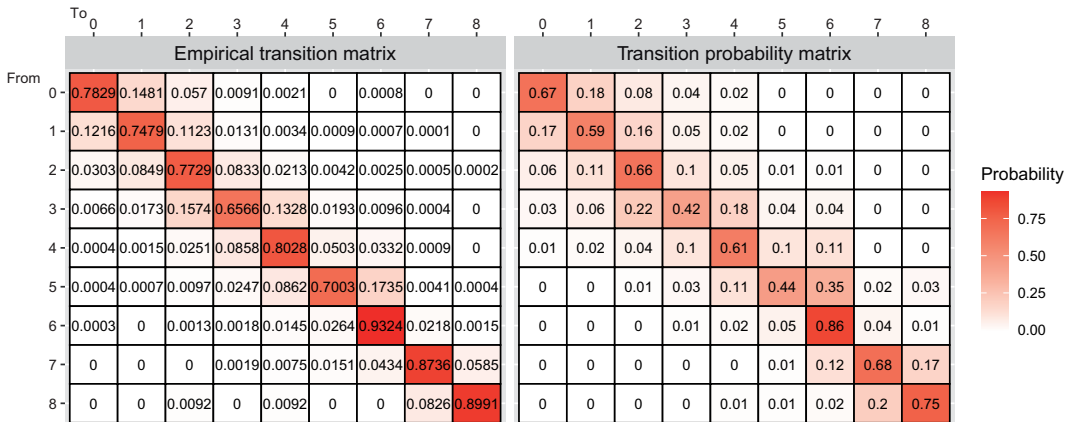


Fig. 3. (Left) normalized empirical transition matrix. (Right) estimated transition probability matrix by SCTMM.

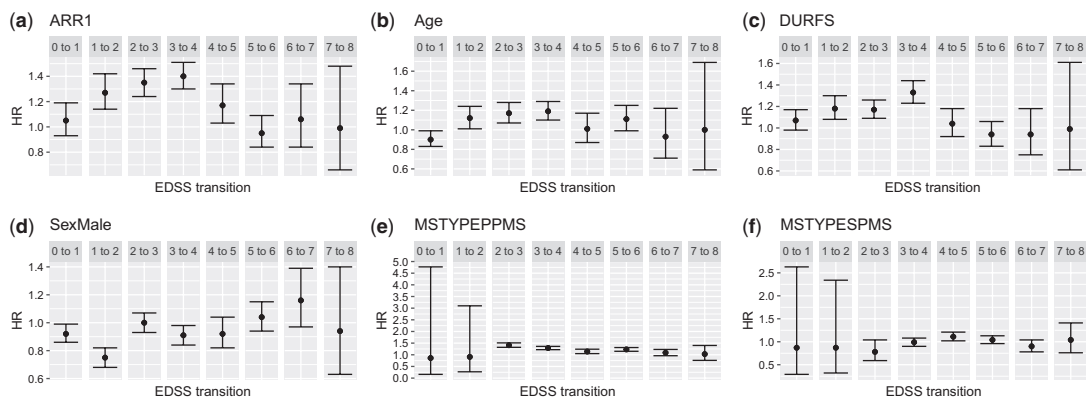


Fig. 4. Estimated hazard ratios of different covariates on successive EDSS state worsening. Error bars show 95% confidence intervals.

for someone with an EDSS of 3, will increase the risk of transition to EDSS 4 by approximately 40% (25%–50%), controlling for other covariates. Figure 4b shows that older age is associated with faster EDSS transition, primarily but not exclusively early in the disease. For instance, having a decade increase in age increases the risk of transitioning from EDSS 2 to 3 by approximately 19% (14%–24%). Figures 4c show some association of duration since the first MS symptoms were observed on the EDSS transitions.

Figure 5 shows a more comprehensive set of heat-map plots of hazard ratios for different covariates, where we do not only show the successive transitions but more EDSS changes up to two jumps of states.

Figures 5a and 5b show that having a relapse in the last year and older age impact the chance of EDSS worsening or improvement particularly at the early stage of the disease (most patients at this stage of the disease are of the ‘relapsing-remitting MS’ sub-type, hence the name). Figure 5c shows the longer the duration of the disease, the higher the chance of further deterioration ( $HR > 1$ ) and

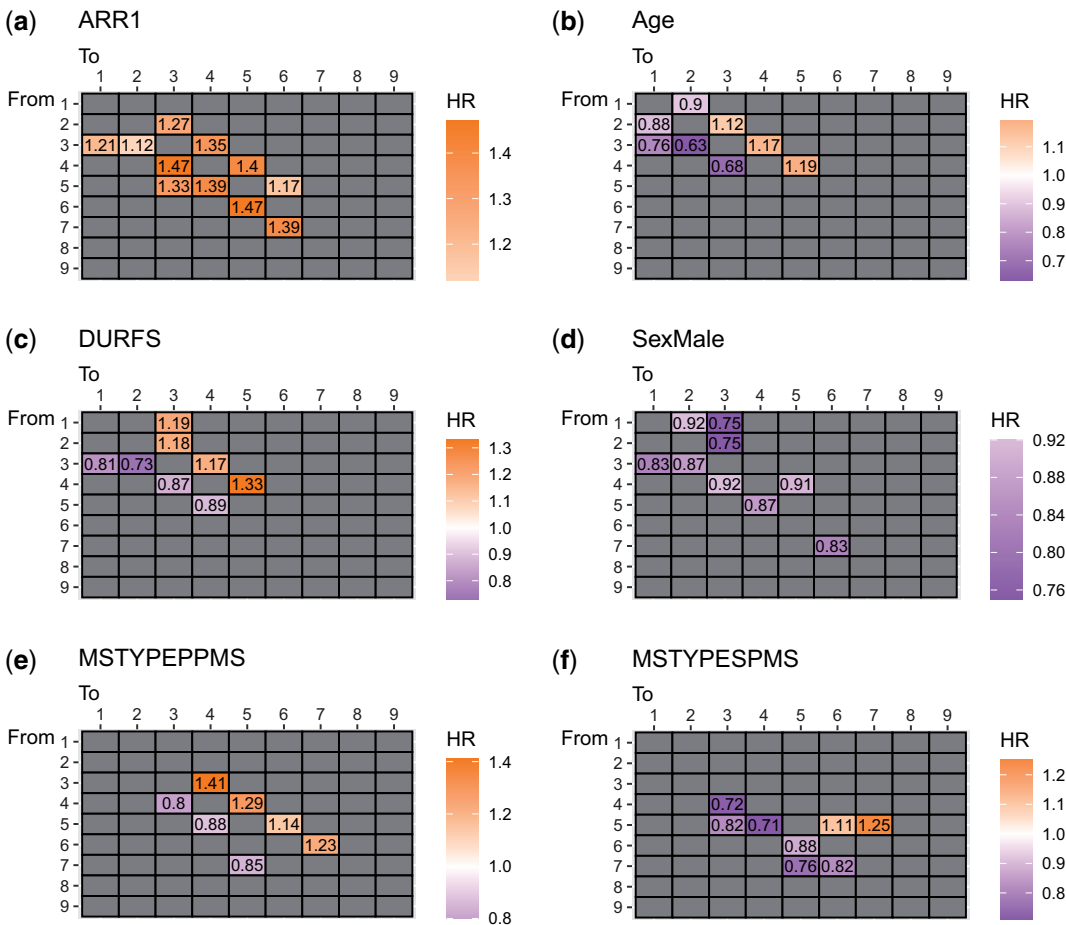


Fig. 5. Estimated significant hazard ratios (HR 95% CI does not include 1) for different covariates on EDSS state transitions.

the lower the chance of recovery ( $HR < 1$ ), controlling for other covariates including age. It may be expected that covariates have an opposing effect on the risk of worsening versus improvement, however, figures 5a and 5d show that this is not the case of having a relapse and being male; i.e. having a relapse in the last year always increase both chances of deterioration and improvement (e.g. recovery from relapses), and males have a lower hazard of transitioning compared to females (this may likely be explained by the higher probability of male patients belonging to the progressive subtypes where the female to male ratio is approximately 1:1 than the relapsing-remitting subtype of MS where the corresponding ratio may be 2:1 or even 3:1). Figures 5e and 5f show that patients with progressive MS (SPMS or PPMS) have a higher chance of deterioration and a reduced chance of recovery compared to patients diagnosed with RRMS.

## 6. CONCLUSION

We have proposed a method to overcome the problems associated with fitting a CTMM with covariates to large-scale data sets. We use a mini-batch stochastic gradient descent algorithm which uses

a random subset of the data set at each iteration, making it practical to fit large scale data. Furthermore, we used the results introduced by Wilcox (1967) and Van Loan (1978) to calculate the derivatives of the matrix exponential, and using this, then proposed a novel approach for computing confidence intervals via two applications of a Padé approximation to find the second derivatives. We also proposed another method for computing confidence intervals based on the approximation of the power series definition of the matrix exponential. The latter is useful when the number of states in the CTMM problem is high ( $\approx S > 20$ ). In a small/mid scale setting, ( $M \leq 1000$ ) our simulation studies show slight out-performance of the proposed SCTMM over the MSM software. In a large-scale setting ( $M \geq 10\,000$ ), where the MSM software is unable to estimate the parameters, the proposed SCTMM can be used and shows a good performance. Some of the important findings in the analysis of NO.MS.2 are as follows:

- The number of relapses in the last year (ARR1) and older age increase the risk of deterioration (EDSS increase) and reduce the chance of recovery (EDSS decrease), particularly in the early stage of the disease.
- Disability accumulation is a slow process in MS: in our large MS data set in which EDSS assessments occur on average approximately every 4 months, most patients stay unchanged in their EDSS state between consecutive visits. The probability of transitioning to a higher/lower EDSS state occurs more frequently in the young and in RRMS compared with the progressive subtypes of the disease (SPMS or PPMS) where patients are more likely to worsen and less likely to recover.
- Our proposed statistical method makes the fitting of Markov models with covariates feasible and scalable to large data sets. It allows the investigation of covariate effects on transition probabilities between (disease) stages, which may find its application far beyond MS.

## 7. SOFTWARE

The code used to implement the approach outlined in the simulation study and data application are available in github at the following link: <https://github.com/farhad-hat/SCTMM>. A sample input data set and complete documentation is available on request from the corresponding author Thomas Nichols ([thomas.nichols@bdi.ox.ac.uk](mailto:thomas.nichols@bdi.ox.ac.uk)).

## ACKNOWLEDGMENTS

The authors acknowledge the work of Jelena Čuklina, Steve Gardiner, and Piet Aarden in coordinating and conducting data wrangling work. They also thank Dieter A. Häring for his detailed comments on the clinical aspects of the article.

*Conflict of Interest:* AO is an employee of Novartis.

## FUNDING

This work was funded in part by Novartis through the Oxford BDI-Novartis Collaboration for AI in Medicine.

## APPENDIX

To compute the first derivatives w.r.t.  $\theta = (q_{ij}^0, \beta_{ij,r})$ , we need to only consider the  $s_{m(k-1)}s_{mk}$  entry of the matrix exponential:

$$\begin{aligned}\partial/\partial q_{ij}^0(I + \mathbf{Q}_{mk}\tau_{mk} + 1/2(\mathbf{Q}_{mk}\tau_{mk})^2) &= \partial/\partial q_{ij}^0(\mathbf{Q}_{mk}\tau_{mk}) \\ &\quad + \partial/\partial q_{ij}^0(1/2(\mathbf{Q}_{mk}\tau_{mk})^2), \\ \partial/\partial \beta_{ij,r}(I + \mathbf{Q}_{mk}\tau_{mk} + 1/2(\mathbf{Q}_{mk}\tau_{mk})^2) &= \partial/\partial \beta_{ij,r}(\mathbf{Q}_{mk}\tau_{mk}) \\ &\quad + \partial/\partial \beta_{ij,r}(1/2(\mathbf{Q}_{mk}\tau_{mk})^2).\end{aligned}\tag{A.1}$$

Now  $\partial/\partial(\cdot)(\mathbf{Q}_{mk}\tau_{mk})$  are as calculated in Section 3, and we are left with calculation of  $\partial/\partial(\cdot)(1/2(\mathbf{Q}_{mk}\tau_{mk})^2)$ . To extract the entry  $s_{m(k-1)}s_{mk}$ , we can use the general form for the second power of the matrix  $\mathbf{Q}_{mk}^2$ ,

$$\begin{aligned}\partial/\partial(\cdot)(1/2(\mathbf{Q}_{mk}\tau_{mk})^2)_{s_{m(k-1)}s_{mk}} &= 1/2\tau_{mk}^2 \partial/\partial(\cdot)(\mathbf{Q}_{mk}^2)_{s_{m(k-1)}s_{mk}} \\ &= 1/2\tau_{mk}^2 \partial/\partial(\cdot)\left(\sum_{n=1}^S q_{s_{m(k-1)}n} q_{ns_{mk}}\right),\end{aligned}\tag{A.2}$$

and the same goes for derivatives w.r.t  $\beta_{ij,r}$ . Now define

$$U = \sum_{n=1}^S q_{s_{m(k-1)}n} q_{ns_{mk}}.\tag{A.3}$$

Applying the chain rule, i.e.,  $\partial(\cdot)/\partial q_{ij}^0 = \partial(\cdot)/\partial q_{ij} \cdot \partial q_{ij}/\partial q_{ij}^0$  (and similarly for  $\beta_{ij,r}$ ), we can consider the three following situations:

1. If  $i = s_{m(k-1)}$  and  $j \neq s_{mk}$ :

$$\begin{aligned}\partial/\partial q_{ij}^0(U) &= \partial/\partial q_{ij}^0\left[\sum_{n=1}^S q_{in} \cdot q_{ns_{mk}}\right] = \partial/\partial q_{ij}^0[q_{ij} \cdot q_{js_{mk}}] = q_{js_{mk}} H_{s_{m(k-1)}j}, \\ \partial^2/\partial q_{ij}^0 \partial q_{ab}^0(U) &= \begin{cases} H_{ab} H_{s_{m(k-1)}j}, & \text{if } a = j \text{ and } b = s_{mk}, \\ 0, & \text{otherwise,} \end{cases} \\ \partial^2/\partial q_{ij}^0 \partial \beta_{ab,c}(U) &= \begin{cases} q_{js_{mk}} z_{mc}(t_{mk}) H_{s_{m(k-1)}j}, & \text{if } a = j \text{ and } b = s_{mk}, \\ q_{js_{mk}} z_{mc}(t_{mk}) H_{s_{m(k-1)}j}, & \text{if } a = s_{m(k-1)} \text{ and } b = j, \end{cases} \\ \partial/\partial \beta_{ij,r}(U) &= q_{ij} z_{mr}(t_{mk}) q_{js_{mk}}, \\ \partial^2/\partial \beta_{ij,r} \partial q_{ab}^0(U) &= \begin{cases} H_{ab} z_{mr}(t_{mk}) q_{js_{mk}}, & \text{if } a = s_{m(k-1)} \text{ and } b = j, \\ q_{s_{m(k-1)}j} z_{mr}(t_{mk}) H_{ab}, & \text{if } a = j \text{ and } b = s_{mk}, \end{cases} \\ \partial^2/\partial \beta_{ij,r} \partial \beta_{ab,c}(U) &= \begin{cases} q_{s_{m(k-1)}j} z_{mr}(t_{mk}) z_{mc}(t_{mk}) q_{js_{mk}}, & \text{if } a = s_{m(k-1)} \text{ and } b = j, \\ q_{s_{m(k-1)}j} z_{mr}(t_{mk}) z_{mc}(t_{mk}) q_{js_{mk}}, & \text{if } a = j \text{ and } b = s_{mk}, \end{cases}\end{aligned}\tag{A.4}$$

where  $H_{ij} = \exp(\sum_{r=1}^R z_{mr}(t_{mk}) \beta_{ij,r})$ .

2. If  $i \neq s_{m(k-1)}$  and  $j = s_{mk}$ :

$$\begin{aligned}
 \partial/\partial q_{ij}^0(U) &= \partial/\partial q_{ij}^0 \left[ \sum_{n=1}^S q_{s_{m(k-1)}n} \cdot q_{nj} \right] = \partial/\partial q_{ij}^0 [q_{s_{m(k-1)}i} \cdot q_{ij}] = q_{s_{m(k-1)}i} H_{is_{mk}}, \\
 \partial^2/\partial q_{ij}^0 \partial q_{ab}^0(U) &= \begin{cases} H_{ab} H_{is_{mk}}, & \text{if } a = s_{m(k-1)} \text{ and } b = i, \\ 0, & \text{otherwise,} \end{cases} \\
 \partial^2/\partial q_{ij}^0 \partial \beta_{ab,c}(U) &= \begin{cases} q_{s_{m(k-1)}i} z_{mc}(t_{mk}) H_{is_{mk}}, & \text{if } a = s_{m(k-1)} \text{ and } b = i, \\ q_{s_{m(k-1)}i} z_{mc}(t_{mk}) H_{is_{mk}}, & \text{if } a = i \text{ and } b = s_{mk}, \end{cases} \\
 \partial/\partial \beta_{ij,r}(U) &= q_{ij} z_{mr}(t_{mk}) q_{s_{m(k-1)}i}, \\
 \partial^2/\partial \beta_{ij,r} \partial q_{ab}^0(U) &= \begin{cases} H_{ab} z_{mr}(t_{mk}) q_{s_{m(k-1)}i}, & \text{if } a = i \text{ and } b = s_{mk}, \\ q_{is_{mk}} z_{mr}(t_{mk}) H_{ab}, & \text{if } a = s_{m(k-1)} \text{ and } b = i, \end{cases} \\
 \partial^2/\partial \beta_{ij,r} \partial \beta_{ab,c}(U) &= \begin{cases} q_{is_{mk}} z_{mr}(t_{mk}) z_{mc}(t_{mk}) q_{s_{m(k-1)}i}, & \text{if } a = i \text{ and } b = s_{mk}, \\ q_{is_{mk}} z_{mr}(t_{mk}) z_{mc}(t_{mk}) q_{s_{m(k-1)}i}, & \text{if } a = s_{m(k-1)} \text{ and } b = i, \end{cases}
 \end{aligned} \tag{A.5}$$

3. If  $i = s_{m(k-1)}$  and  $j = s_{mk}$ :

$$\begin{aligned}
 \partial/\partial q_{ij}^0(U) &= \\
 \partial/\partial q_{ij}^0 \left[ \sum_{n=1}^S q_{in} \cdot q_{nj} \right] &= \partial/\partial q_{ij}^0 [q_{ij} q_{jj} + q_{ii} q_{ij}] = (q_{s_{m(k-1)}s_{m(k-1)}} + q_{s_{mk}s_{mk}}) H_{s_{m(k-1)}s_{mk}}, \\
 \partial^2/\partial q_{ij}^0 \partial q_{ab}^0(U) &= \\
 \begin{cases} H_{s_{m(k-1)}s_{m(k-1)}} H_{s_{m(k-1)}s_{mk}}, & \text{if } a = b = s_{m(k-1)}, \\ H_{s_{mk}s_{mk}} H_{s_{m(k-1)}s_{mk}}, & \text{if } a = b = s_{mk}, \end{cases} \\
 \partial^2/\partial q_{ij}^0 \partial \beta_{ab,c}(U) &= \\
 \begin{cases} z_{mc}(t_{mk}) q_{s_{m(k-1)}s_{m(k-1)}} H_{s_{m(k-1)}s_{mk}}, & \text{if } a = b = s_{m(k-1)}, \\ z_{mc}(t_{mk}) q_{s_{mk}s_{mk}} H_{s_{m(k-1)}s_{mk}}, & \text{if } a = b = s_{mk}, \\ (q_{s_{m(k-1)}s_{m(k-1)}} + q_{s_{mk}s_{mk}}) z_{mc}(t_{mk}) H_{s_{m(k-1)}s_{mk}}, & \text{if } a = s_{m(k-1)} \text{ and } b = s_{mk}, \end{cases} \\
 \partial/\partial \beta_{ij,r}(U) &= \\
 (q_{s_{m(k-1)}s_{m(k-1)}} + q_{s_{mk}s_{mk}}) q_{s_{m(k-1)}s_{mk}} z_{mr}(t_{mk}) H_{s_{m(k-1)}s_{mk}}, \\
 \partial^2/\partial \beta_{ij,r} \partial q_{ab}^0(U) &= \begin{cases} H_{ab} z_{mr}(t_{mk}) q_{s_{m(k-1)}i}, & \text{if } a = i \text{ and } b = s_{mk}, \\ q_{is_{mk}} z_{mr}(t_{mk}) H_{ab}, & \text{if } a = s_{m(k-1)} \text{ and } b = i, \end{cases} \\
 \partial^2/\partial \beta_{ij,r} \partial \beta_{ab,c}(U) &= \\
 \begin{cases} q_{s_{m(k-1)}s_{m(k-1)}} q_{s_{m(k-1)}s_{mk}} z_{mc}(t_{mk}) z_{mr}(t_{mk}) H_{s_{m(k-1)}s_{mk}}, & \text{if } a = b = s_{m(k-1)}, \\ q_{s_{mk}s_{mk}} q_{s_{m(k-1)}s_{mk}} z_{mc}(t_{mk}) z_{mr}(t_{mk}) H_{s_{m(k-1)}s_{mk}}, & \text{if } a = b = s_{mk}, \\ 2((q_{s_{m(k-1)}s_{m(k-1)}} + q_{s_{mk}s_{mk}}) q_{s_{m(k-1)}s_{mk}} z_{mr}(t_{mk}) H_{s_{m(k-1)}s_{mk}}), & \text{if } a = s_{m(k-1)} \text{ and } b = s_{mk}. \end{cases}
 \end{aligned} \tag{A.6}$$

Notice that when  $i \neq s_{m(k-1)}$  and  $j \neq s_{mk}$  then derivatives w.r.t both  $q_{ij}^0$  and  $\beta_{ij,r}$  are zero.

## REFERENCES

- ALVAREZ, J. C., ABE, E., ETTING, I., LE GUEN, M., DEVILLIER, P. AND GRASSIN-DELYLE, S. (2015). Quantification of remifentanyl and propofol in human plasma: a LC-MS/MS assay validated according to the EMA guideline. *Bioanalysis* **7**, 1675–1684.
- COOK, R. J., KALBFLEISCH, J. D. AND YI, G. Y. (2002). A generalized mover–stayer model for panel data. *Biostatistics* **3**, 407–420.
- COX, D. R. AND MILLER, H. D. (1977). *The Theory of Stochastic Processes*. Chapman & Hall.
- DAHLKE, F., ARNOLD, D. L., AARDEN, P., GANJGAHI, H., HÄRING, D. A., ČUKLINA, J., NICHOLS, T. E., GARDINER, S., BERMEL, R. AND WIENDL, H. (2021). Characterisation of MS phenotypes across the age span using a novel data set integrating 34 clinical trials (no. MS cohort): age is a key contributor to presentation. *Multiple Sclerosis Journal* **27**, 2062–2076.
- FLEMING, C. H. AND CALABRESE, J. M. (2021). CTMM: continuous-time movement modeling. R package version 0.6.1.
- FRANCHINI, G., RUGGIERO, V. AND ZANNI, L. (2020). Steplength and mini-batch size selection in stochastic gradient methods. In: Nicosia, G. and others (editors), *International Conference on Machine Learning, Optimization, and Data Science*. Cham: Springer, 259–263.
- HANKS, E. (2018). ctmmove: modeling animal movement with continuous-time discrete-space Markov chains. R package version 1.2.9.
- HSIEH, H.-J., CHEN, T. H.-H. AND CHANG, S.-H. (2002). Assessing chronic disease progression using nonhomogeneous exponential regression Markov models: an illustration using a selective breast cancer screening in Taiwan. *Statistics in Medicine* **21**, 3369–3382.
- JACKSON, C. (2011). Multistate models for panel data: the msm package for R. *Journal of Statistical Software* **38**, 1–28.
- KALBFLEISCH, J. D. AND LAWLESS, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* **80**, 863–871.
- KAPPOS, L., KUHLE, J., MULTANEN, J., KREMENCHUTZKY, M., DI CANTOGNO, E. V., CORNELISSE, P., LEHR, L., CASSET-SEMANAZ, F., ISSARD, D. AND UITDEHAAG, B. M. J. (2015). Factors influencing long-term outcomes in relapsing–remitting multiple sclerosis: Prisms-15. *Journal of Neurology, Neurosurgery & Psychiatry* **86**, 1202–1207.
- KURTZKE, J. F. (1983). Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology* **33**, 1444.
- LUBLIN, F. D., HÄRING, D. A., GANJGAHI, H., OCAMPO, A., HATAMI, F., ČUKLINA, J., AARDEN, P., DAHLKE, F., ARNOLD, D. L., WIENDL, H. and others (2022). How patients with multiple sclerosis acquire disability. *Brain* **145**, 3147–3161.
- LUO, Y., STEPHENS, D. A., VERMA, A. AND BUCKERIDGE, D. L. (2021). Bayesian latent multistate modeling for nonequidistant longitudinal electronic health records. *Biometrics* **77**, 78–90.
- MANDEL, M., GAUTHIER, S. A., GUTTMANN, C. R. G., WEINER, H. L. AND BETENSKY, R. A. (2007). Estimating time to event from longitudinal categorical data: an analysis of multiple sclerosis progression. *Journal of the American Statistical Association* **102**, 1254–1266.
- MANDEL, M., MERCIER, F., ECKERT, B., CHIN, P. AND BETENSKY, R. A. (2013). Estimating time to disease progression comparing transition models and survival methods—an analysis of multiple sclerosis data. *Biometrics* **69**, 225–234.
- MARSHALL, G. AND JONES, R. H. (1995). Multistate models and diabetic retinopathy. *Statistics in Medicine* **14**, 1975–1983.



- MOLER, C. AND VAN LOAN, C. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review* **45**, 3–49.
- PERRONE, M. P., KHAN, H., KIM, C., KYRILLIDIS, A., QUINN, J. AND SALAPURA, V. (2019). Optimal mini-batch size selection for fast gradient descent. arXiv, arXiv:1911.06459, preprint: not peer reviewed.
- R CORE TEAM. (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- RUNMARKER, B. AND ANDERSEN, O. (1993). Prognostic factors in a multiple sclerosis incidence cohort with twenty-five years of follow-up. *Brain* **116**, 117–134.
- SAKRISON, D. J. (1965). Efficient recursive estimation; application to estimating the parameters of a covariance function. *International Journal of Engineering Science* **3**, 461–483.
- VAN LOAN, C. (1978). Computing integrals involving the matrix exponential. *IEEE Transactions on Automatic Control* **23**, 395–404.
- VUKUSIC, S. AND CONFAYREUX, C. (2003). Prognostic factors for progression of disability in the secondary progressive phase of multiple sclerosis. *Journal of the Neurological Sciences* **206**, 135–137.
- VUKUSIC, S. AND CONFAYREUX, C. (2007). Natural history of multiple sclerosis: risk factors and prognostic indicators. *Current Opinion in Neurology* **20**, 269–274.
- WILCOX, R. M. (1967). Exponential operators and parameter differentiation in quantum physics. *Journal of Mathematical Physics* **8**, 962–982.
- WILLIAMS, J. P., STORLIE, C. B., THERNEAU, T. M., JACK, JR, C. R. AND HANNIG, J. (2020). A Bayesian approach to multistate hidden Markov models: application to dementia progression. *Journal of the American Statistical Association* **115**, 16–31.
- ZURAWSKI, J., GLANZ, B. I., CHUA, A., LOKHANDE, H., ROTSTEIN, D., WEINER, H., ENGLER, D., CHITNIS, T. AND HEALY, B. C. (2019). Time between expanded disability status scale (EDSS) scores. *Multiple Sclerosis and Related Disorders* **30**, 98–103.

[Received September 8, 2022; revised May 26, 2023; accepted for publication May 30, 2023]