Risk factor identification in heterogeneous disease progression with L1-regularized multi-state models

Xuan Dang · Shuai Huang · Xiaoning Qian

Received: date / Accepted: date

Abstract Multi-state model (MSM) is a useful tool to analyze longitudinal data for modeling disease progression at multiple time points. While the regularization approaches to variable selection have been widely used, extending them to MSM remains largely unexplored. In this paper, we have developed the L1-regularized multi-state model (L1MSTATE) framework that enables parameter estimation and variable selection simultaneously. The regularized optimization problem was solved by deriving a one-step coordinate descent algorithm with great computational efficiency. The L1MSTATE approach was evaluated using extensive simulation studies, and it showed that L1MSTATE outperformed existing regularized multi-state models in terms of the accurate identification of risk factors. It also outperformed the un-regularized multi-state models (MSTATE) in terms of identifying the important risk factors in situations with small sample sizes. The power of L1MSTATE in predicting the transition probabilities comparing with MSTATE was demonstrated using the Europe Blood and Marrow Transplantation (EBMT) dataset. The L1MSTATE was implemented in the open-access R package '**L1mstate**'.

Keywords Multi-state model · L1-regularization · Longitudinal data · Rare transition prediction · Variable selection.

Xuan Dang
E-mail: xuandt89@tamu.edu
Texas A&M University, College Station, TX 77840, USA
Shuai Huang
E-mail: shuaih@uw.edu
University of Washington, Seattle, WA 98195, USA
Xiaoning Qian
E-mail: xqian@ece.tamu.edu
Texas A&M University, College Station, TX 77840, USA

1 Introduction

Multi-state model (MSM) has been one of effective methods for disease modeling, and it has been applied to studying liver cancer [27], breast cancer [36][15][10], abdominal aortic aneurysms [24], heart transplantation [40][30], HIV infection and AIDS [31][19], Alzheimer disease [11], diabetic complication [33][5], cervical cancer [29], and liver cirrhosis [7], just to name a few. It can model patient's disease development trajectory across a series of transitions between various stages or states, under influence of some risk factors. First, it allows researchers to make an assessment about how the risk factors exert different effects on different stages of the process and how the risk factors influence on different transitions of the process. Second, it enables researchers to obtain more accurate predictions of transition probabilities.

In this paper, we adopted the MSM framework by specifying the transition-specific hazard models. Our main objective is to identify the risk factors associated with the transition hazard rates of disease progression. Although non-parametric transition hazard models do not impose any constraint and may be more flexible, it is used more often to estimate the cumulative transition hazard rates than the transition hazard rates [1]. Semi-parametric transition hazard models that do not require to specify the transition-specific baseline hazard functions are more suitable for our purpose. Specifically, the Cox's proportional hazards model was used for the transition-specific hazard rates to incorporate risk factors into multi-state models. The multi-state model parameters were estimated by maximizing the likelihood function that was formulated using the counting process [6]. The transition-specific baseline hazards were assumed to be the same for all individuals but vary over time, allowing us to construct the partial likelihood function that reduces computation burden but still makes good estimations of parameters [25]. Regarding the censored data, we focused on two types of censoring data: right-censored and left-truncation data.

Currently, the multistate models lack an efficient and practical variable selection method to identify the risk factors associated with the transition hazard rates. Let us consider a MSM with the number of the risk factors is P and the number of transitions between the stages is Q. Then, there are 2^{PQ} possible models to consider if using stepwise forward selection [33] method. Hence, such kinds of variable selection methods are suitable when the number of risk factors and the number of transitions is relatively small. However, in modern applications, both P and Q increase dramatically with our increasing data collection capacity. They result in complicated optimization problems which are challenging to compute, and they can lead unstable estimates of parameters. In addition, in many studies, especially in medical research, there is a limited number of observations given the number of parameters in complex multi-state models. In this paper, the regularization approaches have been used to address these challenges. Intuitively, these approaches incorporate the prior knowledge about sparse structures of multi-state models using the sparse-inducing penalties, which results in better parameter estimations and allows variable selection simultaneously.

Even though the regularization methods are increasingly popular in statistics and machine learning very little has extended to MSMs. The current literature on this subject shows there are two works that have been published in this direction. The first one by Huang *et al.* 2018 [23] presented a regularized continuous-time Markov model with the elastic net penalty. The transition hazard rates were specified as constant over time. In addition, their method relied on a method developed by [26]: it estimated the transition rates from the transition probabilities of the discrete-time Markov chain embedded in the Markov process (embedded Markov chain). It does not derive the transition rates from event (state) counts and transitions since the transition times are not observed. In other words, it does not follow the counting process perspective. Therefore, their work is different from ours in scope and methodology.

The second one from Reulen *et al.* 2016 [38] did variable selection by imposing the fused-lasso penalties including L1-penalties of transition-specific risk factor coefficients and their differences between transitions. In this paper, we propose the L1-penalties of transition-specific risk factor coefficients that are similar to the fused-lasso approach in [38], in which cross-transition effects are explicitly modeled by introducing the fused penalties. The difference of our implementation from [38] is, instead of adopting the penalized iteratively re-weighted least squares (PIRLS) algorithm presented in Oelker *et al.* 2017 [35] for model inference, we have derived a cyclical one-step coordinate descent algorithm to solve the optimization problem with exact L1-penalties. In addition to potential problems of not having exact zero model coefficients due to the approximation of L1-penalties, PIRLS is a second-order optimization algorithm that has high computation cost and potential convergence problems [35]. Our optimization algorithm in this paper solves for exactly L1-penalties resulting in fewer nonzero coefficients for variable selection, with high efficiency in computation and significant reduction in memory usage.

Another common problem in many studies is that multi-state models include some *rare* transitions that have relatively small number of observations. In such cases, the traditional (un-regularized) multi-state model approach tends to produce the inaccurate predictions of the probabilities of *rare* transitions. In this paper, we demonstrated that the L1-regularized multi-state models can be used to alleviate this problem, and thus produce better predictions of the transition probabilities.

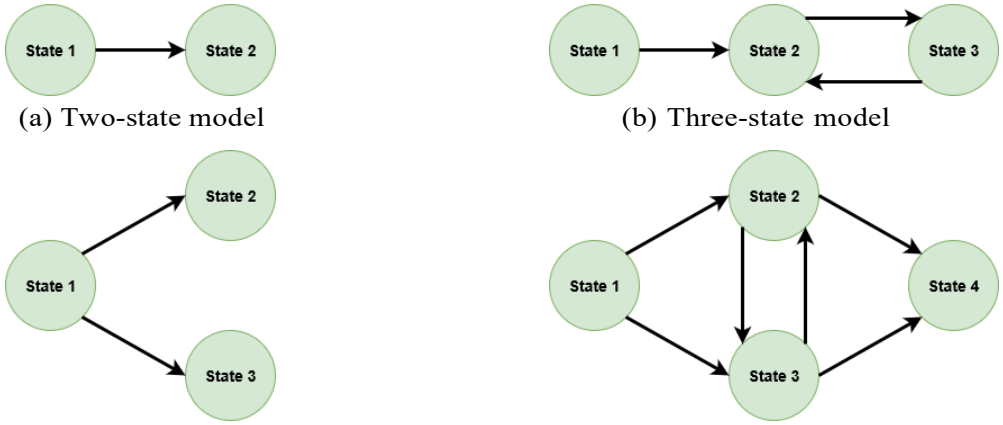
The rest of the article is organized as follows. In Section 2, we reviewed critical details of the multi-state models, including its formulation and the partial likelihood function of the multi-state models. In Section 3, we introduced our formulation of the L1-regularized partial likelihood function of the multi-state models and the algorithms to solve the corresponding optimization problems. We presented the main formulae to predict the

transition probabilities. In Section 4, we compared the performance of our method via simulation studies. We demonstrated the prediction power of our method using a real data. Discussion was presented in Section 6. Lastly, we ended with conclusions and future works in Section 7.

2 Review of Multi-state Models (MSMs)

2.1 Formulation of the multi-state models

Multi-state models compose of multiple states and transitions between the states under influence of risk factors. Figure 1 depicts some examples of the multi-state models in characterizing a variety of situations with different number of states and transition structures between the states. For example, in Figure 1.c, there are three states. The arrows illustrate the clinically eligible transitions between the states. The state to which the individual is going to move, and the time of this change, is impacted by the transition intensities (so-called hazard rates) that represent the instantaneous risk of moving from one state to another. These hazard rates may also depend on individual-specific risk factors. In our paper, we assume that the risk factors are constant over time. The states and structure of the transitions are usually pre-defined based on domain knowledge of the disease. The main statistical task is to estimate the transition intensities between states and their relationships with the risk factors.



(a) Two-state model (b) Three-state model

(c) Three-state model (d) Four-state model

Fig. 1: Some multi-state models. *Note*: Arrows show the clinically eligible transitions for each multi-state model.

We specify the transition-specific hazard rates $\alpha_{hj}(t)$ using Cox proportional hazards model [12] with the transition-specific baseline hazard rates $\alpha_{hj}^{(0)}(t)$ and time-fixed risk factors $\boldsymbol{X}$:

$$\alpha_{hj}(t) = \alpha_{hj}^{(0)}(t)\exp(\boldsymbol{\beta}_{hj}^{T}\boldsymbol{X}) \qquad (1)$$

where $\boldsymbol{X} = (x_1, x_2, \cdots, x_P)^T$ is an $P-$dimensional vector of time-fixed risk factors; $\boldsymbol{\beta}_{hj}^{T}$ is a $P-$dimensional vector of time-fixed coefficients.

2.2 Likelihood function of the multi-state models

Then, we can derive the likelihood formulation of the multi-state model. Consider M individuals, $S_i(t)$ is the observed multi-state model for the i^{th} individual over interval $[0, \tau_i]$, where τ_i is a fixed time of termination of observation for individual i. Denote $N_{hj}^{i}(t)$ be the number of allowed transitions $h \to j$ of the i^{th} individual during $[0, t]$, and $\alpha_{hj}^{i}(t)$ be transition intensities or transition-specific hazard rates of the i^{th} individual. The transition times T_{hj}^{ik} can be described as $0 < T_{hj}^{i1}(t) < \cdots < T_{hj}^{iN_{hj}^{i}(\tau_i)}(t) \leq \tau_i$, where $k \in \{1, \ldots, N_{hj}^{i}(\tau_i)\}$. The full likelihood function could be derived as

$$L = \prod_{i=1}^{M}\prod_{j\neq h}\prod_{k=1}^{N_{hj}^{i}(\tau_i)}\left[\alpha_{hj}^{i}(T_{hj}^{ik})\exp\left(-\int_{0}^{T_{hj}^{ik}}\alpha_{hj}^{i}(t)dt\right)\right]$$

Assume that individual-specific risk factors are constant over time, the transition-specific hazard rate $\alpha_{hj}^{i}(t)$ each individual i can be written as Eq. (1). The full likelihood function becomes

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{M}\prod_{j\neq h}\prod_{k=1}^{N_{hj}^{i}(\tau_i)}\left[\alpha_{hj}^{i(0)}(t)\exp(\boldsymbol{\beta}_{hj}^{T}\boldsymbol{X}^{i})\exp\left(-\int_{0}^{T_{hj}^{ik}}\alpha_{hj}^{i(0)}(t)\exp(\boldsymbol{\beta}_{hj}^{T}\boldsymbol{X}^{i})dt\right)\right] \qquad (2)$$

where $X^i = \left(x_1^i, x_2^i, \ldots, x_P^i\right)^T$ is $P-$ dimensional vector of time-constant risk factors for the i^{th} individual.

2.3 Partial likelihood function for multi-state model

Instead of using the above full likelihood function, we used the partial likelihood function. More details can be found in Andersen *et al.* 1993 [6]. It only keeps the terms that contain all the information about β and gets rid of the terms that contain the information about the baseline hazard. This achieves computational efficiency and still makes good inference for β.

Let $Y_{hj}^{ik}(t) = 1_{\{t \le T_{hj}^{ik}\}}$, i.e., in this definition $Y_{hj}^{ik}(T_{hj}^{ik})$ indicates that the i^{th} individual at risk in transition from state h to state j at time T_{hj}^{ik}. Assume that the transition-specific baseline hazards are the same for all individuals but can vary freely with time, i.e., $\alpha_{hj}^{i(0)}(t) = \alpha_{hj}^{(0)}$. The partial likelihood function of the multi-state model that will be used in our paper

$$L^p(\boldsymbol{\beta}) = \prod_{j \neq h} \prod_{i=1}^{M} \prod_{k=1}^{N_{hj}^i(\tau_i)} \frac{\exp(\boldsymbol{\beta}_{hj}^T \boldsymbol{X}^i)}{\sum_{i=1}^{M} \sum_{k=1}^{N_{hj}^i(\tau_i)} \exp(\boldsymbol{\beta}_{hj}^T \boldsymbol{X}^i) Y_{hj}^{ik}(t)},$$

Its negative log-partial likelihood function is derived as

$$l(\boldsymbol{\beta}) = -\log(L^p(\boldsymbol{\beta})) = -\sum_{j \neq h} \sum_{i=1}^{M} \sum_{k=1}^{N_{hj}^i(\tau_i)} \left[(\boldsymbol{\beta}_{hj}^T \boldsymbol{X}^i) - \log\left(\sum_{i=1}^{M} \sum_{k=1}^{N_{hj}^i(\tau_i)} \exp(\boldsymbol{\beta}_{hj}^T \boldsymbol{X}^i) Y_{hj}^{ik}(t) \right) \right] \quad (3)$$

2.4 Data structure for parameter estimation by partial likelihood maximization

We follow the data structure described in Putter *et al.* 2007 [37]. One example as shown in Table 1 was collected in deWreede *et al.* 2010 [14]. In this format, each individual has many rows as the number of transitions for which she/he is at risk. Particularly, in Table 1, each row shows one transition of each individual that is composed by state_{from} and state_{to}. The corresponding times for state_{from} and state_{to} are time_{start} and time_{stop}. The difference between time_{start} and time_{stop} measures the transition times that represent the duration for which individual is at risk. The censoring information is captured by a transition-specific censoring indicator $\boldsymbol{\delta}_{status}$. For example, patient 1 contributes two lines of data for the period: start at $t = 0$ and stop at $t = 151$. She/he started at state 2 and was at risk to transfer to state 1 and state 3. The recorded $\boldsymbol{\delta}_{status}$ of transition $2 -> 1$ was 0, which indicates that the event (transition) time was censored, while the recorded $\boldsymbol{\delta}_{status}$ of transition $2 -> 3$ was 1, which indicates that the event time was observed.

Table 1: Example of long-format data

	Patient id	state_{from}	state_{to}	transition	time_{start}	time_{stop}	δ_{status}	treatment
1	1	2	1	3	0	151	0	Placebo
2	1	2	3	4	0	151	1	Placebo
3	2	2	1	3	0	251	1	Placebo
4	2	2	3	4	0	251	0	Placebo
5	2	1	2	1	251	434	1	Placebo
6	2	1	3	2	251	434	0	Placebo
7	2	2	1	3	434	729	1	Placebo
8	2	2	3	4	434	729	0	Placebo
9	2	1	2	1	729	1735	1	Placebo
10	2	1	3	2	729	1735	0	Placebo
11	2	2	1	3	1735	2088	1	Placebo
12	2	2	3	4	1735	2088	0	Placebo
13	2	1	2	1	2088	2467	0	Placebo
14	2	1	3	2	2088	2467	1	Placebo

Following this data structure, suppose that there are in total Q observable transition types. Assume that the dataset has N rows, and denote N_q be the number of rows for transition q, it is easy to see that $N = \sum_q N_q$. With a slight abuse of notation, $\boldsymbol{X}_q$ is the $N_q \times P$ risk factors matrix corresponding to q−transition; X_q^i is the P −dimensional column-vector where $q = 1, 2, \ldots, Q$ and $i = 1, 2, \ldots, N_q$. The formulation of the negative log-partial likelihood function in Eq. (3) could be rewritten as

$$l(\beta) = \sum_q l_q(\beta_q) \quad (4)$$

where

$$l_q(\beta_q) = -\sum_{i=1}^{D_q} \left[(\beta_q^T \boldsymbol{X}_q^i) - \log\left(\sum_{n=1}^{N_q} \exp(\beta_q^T \boldsymbol{X}_q^i) Y_q^n(t_i) \right) \right] = -\sum_{i=1}^{D_q} \left[(\beta_q^T \boldsymbol{X}_q^i) - \log\left(\sum_{r \in R_q^i} \exp(\beta_q^T \boldsymbol{X}_q^r) \right) \right] \quad (5)$$

where D_q is the set of indices of the exact transition times for the transition type q, $Y_q^n(t) = 1_{\{t_q^n \ge t_i\}}$ indicates whether the nth individual is at risk to transition q just before time ti, and $R_q^i = \sum_n Y_q^n(t_i) = \sum_n 1_{\{t_q^n \ge t_i\}}$ is a set of indices r that comprised of all individuals observed to be at risk to transition q with times $\ge t_i$.

Remark: As shown in above, we use only information about the observed states at a set of times when we assume that the distribution of transition times provides no information about the distribution of censorship times and

vice versa. It is so-called the independent censoring [6]. We also assume that the observation time is the exact transition time and there are no transitions between the observation times for each individual. With the formulation of the negative log-partial-likelihood function in Eq. (4), two kinds of incomplete observations are particularly tractable [8]: right-censoring and left-truncation. Note that if the individual is observed from the beginning (i.e., the first state, such as healthy) to the end (i.e., the final state, such as death), then the whole trajectory of the process has been observed and it is called complete observation. Otherwise, right-censoring means that the individual is observed from the beginning to a certain time that has not reached the final state. Left-truncation means that the process has not been observed from the beginning, rather, the observation happens in the middle of the trajectory of the transitions.

3 L1-Regularized multi-state model (L1MSTATE)

3.1 Partial likelihood formulation for L1MSTATE

By minimizing the negative log-partial likelihood formulated in Eq. (4), we can estimate the parameters of a multi-state model, i.e., the coefficients $\boldsymbol{\beta}$. As existing methods could not scale up to high-dimensional applications when there are a large number of risk factors and a large number of transitions, in this paper, we propose a L1-regularized partial likelihood formulation for MSM following the framework as the least absolute shrinkage and selection operator (LASSO) [43]. This leads to the following formulation:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} && l(\beta) \\ &\text{subject to} && \sum_q \sum_p |\beta_q^p| \le C, \end{aligned} \tag{6}$$

where $q = 1, 2, \ldots, Q; p = 1, 2, \ldots, P; C > 0$. Recall that, Q is the number of observable transitions, and P is the number of risk factors. This minimization problem is equivalent to minimizing the problem given by the Lagrangian formulation:

$$\sum_q \frac{1}{N_q} l_q(\beta_q) + \lambda \Big(\sum_q \sum_p |\beta_q^p| \Big),$$

with respect to $\boldsymbol{\beta}$. Different weights are assigned to transitions using factors N_q where $q = 1, 2, \ldots, Q$. It is similar to assign different shrinkage parameters per transition. Intuitively, the rare transitions are shrunk more than for common transitions. Our formulation in Eq. (6) could be reformulated as

$$\hat{\beta} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[\sum_q \frac{1}{N_q} l_q(\beta_q) + \lambda \Big(\sum_q \sum_p |\beta_q^p| \Big) \right]. \tag{7}$$

3.2 Computational algorithm for solving Eq. (7)

The transition-specific negative log-partial-likelihood function $l_q(\beta_q)$ is smooth with respect to β_q so that its first two partial derivatives are continuous. Thus, $l_q(\beta_q)$ can be locally approximated by

$$l_q(\beta_q) \approx l_q(\tilde{\beta}_q) + (\beta_q - \tilde{\beta}_q)^T l_q'(\tilde{\beta}_q) + \frac{1}{2}(\beta_q - \tilde{\beta}_q)^T l_q''(\tilde{\beta}_q)(\beta_q - \tilde{\beta}_q), \tag{8}$$

where

$$l_q'(\tilde{\beta}_q) = \frac{\partial l_q}{\partial \beta_q}(\tilde{\beta}_q) \text{ and } l_q''(\tilde{\beta}_q) = \frac{\partial^2 l_q}{\partial \beta_q \partial \beta_q^T}(\tilde{\beta}_q),$$

The transition-specific linear predictor, $\eta_q = X_q \beta_q$, includes D_q elements $\eta_q^i = \beta_q^T X_q^i$, where $i = 1, \ldots, D_q$. Plugging them in Eq. (5) and Eq. (8), we have the transition-specific negative log-partial likelihood function

$$l_q(\boldsymbol{\eta}_q) = -\sum_{i=1}^{D_q} \Big[\boldsymbol{\eta}_q^i - \log\Big(\sum_{r \in R_q^i} \exp(\boldsymbol{\eta}_q^r) \Big) \Big]$$

Its approximated form is

$$l_q(\boldsymbol{\eta}_q) \approx \frac{1}{2} \Big(\boldsymbol{\eta}_q - z(\tilde{\eta}_q) \Big)^T l_q''(\tilde{\boldsymbol{\eta}}_q) \Big(\boldsymbol{\eta}_q - z(\tilde{\eta}_q) \Big),$$

with

$$z(\tilde{\eta}_q) = \tilde{\eta}_q - \Big(l_q''(\tilde{\boldsymbol{\eta}}_q) \Big)^{-1} l_q'(\tilde{\eta}_q);\ l_q'(\tilde{\eta}_q) = \frac{\partial l_q}{\partial \boldsymbol{\eta}_q}(\tilde{\eta}_q);\ l_q''(\tilde{\eta}_q) = \frac{\partial^2 l_q}{\partial \boldsymbol{\eta}_q \partial \boldsymbol{\eta}_q^T}(\tilde{\eta}_q),$$

Hastie and Tibshirani (1990, Chapter 8) [21] suggested to replace $l_q''(\tilde{\eta}_q)$ by a diagonal matrix D with the diagonal elements of $l_q''(\tilde{\eta}_q)$, because the optimal βq will not change when the off-diagonal elements of $l_q''(\tilde{\eta}_q)$ are smaller than the diagonal elements. This will greatly alleviate our analytic efforts since we only need to compute the first

order derivative $l_q'(\tilde{\eta}_q)$ and the diagonal entry of the second order derivative $l_q''(\tilde{\eta}_q)$; $l_q'(\tilde{\eta}_q)$ is a vector with elements $\left(l_q'(\tilde{\eta}_q)\right)_d$ that could be derived as

$$l_q^{'}(\tilde{\boldsymbol{\eta}}_q)_d = \frac{\partial l_q(\boldsymbol{\eta}_q)}{\partial \eta_q^d} = -\delta_d + \sum_{i=1}^{D_q} \frac{\sum_{d\in R_q^i} \exp(\eta_q^d)}{\sum_{r\in R_q^i} \exp(\eta_q^r)} = -\delta_d + \sum_{i=1}^{D_q} \sum_{d\in R_q^i} \frac{\exp(\eta_q^d)}{\sum_{r\in R_q^i} \exp(\eta_q^r)} = -\delta_d + \sum_{i\in C_q^d} \frac{\exp(\eta_q^d)}{\sum_{r\in R_q^i} \exp(\eta_q^r)} \quad (9)$$

where $d = 1, 2, \ldots, N_q$, and C_q^d is the $q-$transition set of i with $t_d \geq t_i$. The diagonal entry of $l_q''(\tilde{\eta}_q)$ could be derived as

$$l_q^{''}(\tilde{\boldsymbol{\eta}}_q)_{d,d} = \frac{\partial}{\partial \eta_q^d}\left(\frac{\partial l_q(\eta_q)}{\partial \eta_q^d}\right) = \sum_{i\in C_q^d} \left[\frac{\exp(\eta_q^d)}{\sum_{r\in R_q^i} \exp(\eta_q^r)} - \frac{(\exp(\eta_q^d))^2}{\left(\sum_{r\in R_q^i} \exp(\eta_q^r)\right)^2}\right] \quad (10)$$

Let

$$M(\boldsymbol{\beta}_q) = \frac{1}{2N_q}\left(\boldsymbol{\eta}_q - \boldsymbol{z}(\tilde{\boldsymbol{\eta}}_q)\right)^T l_q^{''}(\tilde{\boldsymbol{\eta}}_q)\left(\boldsymbol{\eta}_q - \boldsymbol{z}(\tilde{\boldsymbol{\eta}}_q)\right) + \lambda\left(\sum_p |\beta_q^p|\right)$$

The training algorithm for L1MSTATE is shown in the pseudo code in Algorithm 1. The remaining task is to solve the optimization problem in Eq. (11):

$$\hat{\boldsymbol{\beta}}_q = \underset{\boldsymbol{\beta}_q}{\operatorname{argmin}}\, M(\boldsymbol{\beta}_q), \quad (11)$$

Let w_q be the N_q $-$dimensional vector of diagonal entries of matrix D. We rewrite $M(\boldsymbol{\beta}_q)$ as

$$M(\boldsymbol{\beta}_q) = \frac{1}{2N_q}\sum_{i=1}^{N_q}\left[\boldsymbol{w}_q^i\left(\boldsymbol{z}(\tilde{\boldsymbol{\eta}}_q)_i - \sum_{p\neq g} X_{q,p}^i \beta_q^p - X_{q,g}^i \beta_q^g\right)^2\right] + \lambda\left(\sum_p |\beta_q^p|\right)$$

Hence, Eq. (11) becomes

$$\hat{\boldsymbol{\beta}}_q = \underset{\boldsymbol{\beta}_q}{\operatorname{argmin}}\, \frac{1}{2N_q}\sum_{i=1}^{N_q}\left[\boldsymbol{w}_q^i\left(\boldsymbol{z}(\tilde{\boldsymbol{\eta}}_q)_i - \sum_{p\neq g} X_{q,p}^i \beta_q^p - X_{q,g}^i \beta_q^g\right)^2\right] + \lambda\left(\sum_p |\beta_q^p|\right) \quad (12)$$

The coordinate descent algorithm is used to solve Eq. (12). In particular, we derive the one-step coordinate descent algorithm that updates one element at each iteration with all the other elements fixed to the latest value.

Algorithm 1: Pseudocode for L1-penalized multi-state model (L1MSTATE)

Result: $\hat{\beta}$
Input: Long-format data described in Section 2.4;
while $(q > 0$ *and* $q \leq Q)$ **do**
  Compute $\tilde{\eta}_q = \mathbf{X}_q\tilde{\beta}_q$; $l_q^{'}(\tilde{\eta}_q)$; $l_q^{''}(\tilde{\eta}_q)$; $\mathbf{z}(\tilde{\eta}_q) = \tilde{\eta}_q - l_q^{''}(\tilde{\eta}_q)^{-1} l_q^{'}(\tilde{\eta}_q)$
  Find $\hat{\beta}_q = \underset{\boldsymbol{\beta}_q}{\operatorname{argmin}}\, M(\beta_q)$; Update $\tilde{\beta}_q = \hat{\beta}_q$
end

Specifically, for instance, while the current step focuses on β_q^p with given estimates for β_q^p for all $p \neq g$, we compute the first order derivative of $M(\beta_q)$ as follows

$$\frac{\partial M(\boldsymbol{\beta}_q)}{\partial \beta_q^g} = \frac{1}{N_q}\sum_{i=1}^{N_q}\left[\boldsymbol{w}_q^i\left(\boldsymbol{z}(\tilde{\boldsymbol{\eta}}_q)_i - X_q^i\boldsymbol{\beta}_q\right)(-X_{q,g}^i)\right] + \lambda \mathrm{sgn}(\beta_q^g) \quad (13)$$

where with $g = 1, 2, \ldots, P$

$$\mathrm{sgn}(\beta_q^g) = \begin{cases} 1, & \text{if } \beta_q^g > 0 \\ -1, & \text{if } \beta_q^g < 0 \\ [-1, 1], & \text{otherwise.} \end{cases}$$

Solving Eq. (13) yields the soft-thresholding rule that is

$$\hat{\beta}_q^g = \frac{f\left(\frac{1}{N_q}\sum_{i=1}^{N_q}\left[\boldsymbol{w}_q^i X_{q,g}^i\left(\boldsymbol{z}(\tilde{\boldsymbol{\eta}}_q)_i - \sum_{p\neq g} X_{q,p}^i \beta_q^p\right)\right], \lambda\right)}{\frac{1}{N_q}\sum_{i=1}^{N_q} w_q(\tilde{\boldsymbol{\eta}}_q)_i (X_{q,g}^i)^2} \quad (14)$$

where

$$f(x,\lambda)=\operatorname{sgn}(x)(|x|-\lambda)=\begin{cases}x-\lambda, & \text{if } x>0 \text{ and } |x|>\lambda\\ x+\lambda, & \text{if } x<0 \text{ and } |x|>\lambda\\ 0, & \text{if } |x|\le\lambda.\end{cases}$$

Note that the first term in the numerator can be derived by using Equations. (9) and (10):

$$\boldsymbol{w}_q^i X_{q,g}^i\left(\mathbf{z}(\tilde{\boldsymbol{\eta}}_q)_i-\sum_{p\neq g}X_{q,p}^i\beta_q^p\right)=\tilde{\beta}_q^g\boldsymbol{w}_q(\boldsymbol{X}_{q,g})^2-l_q'(\tilde{\boldsymbol{\eta}}_q)\boldsymbol{X}_{q,g},$$

So, we have a simple form of estimated coefficient as follows

$$\hat{\beta}_q^g=\frac{f\left(\frac{1}{N_q}\left[\tilde{\beta}_q^g\boldsymbol{w}_q(\boldsymbol{X}_{q,g})^2-l_q'(\tilde{\boldsymbol{\eta}}_q)\boldsymbol{X}_{q,g}\right],\lambda\right)}{\frac{1}{N_q}\boldsymbol{w}_q(\boldsymbol{X}_{q,g})^2}. \qquad (15)$$

It is worthy of mentioning that the solution for LASSO depends on the scales of risk factors [22]. A frequently used method to solve this problem is to standardize the risk factors first. The estimated coefficients of the risk factors can always be transformed back to the original scales for the sake of interpretation. The one-step coordinate descent is summarized in Algorithm 2.

Algorithm 2: One step coordinate descent algorithm for L1-penalized multi-state model (L1MSTATE).

Result: $\hat{\beta}$
Input: Long-format data described in Section 2.4;
while *($q>0$ and $q\le Q$)* **do**
  Compute $\tilde{\boldsymbol{\eta}}_q=\mathbf{X}_q\tilde{\boldsymbol{\beta}}_q$; $l_q'(\tilde{\boldsymbol{\eta}}_q)$; $l_q''(\tilde{\boldsymbol{\eta}}_q)$; $\mathbf{z}(\tilde{\boldsymbol{\eta}}_q)=\tilde{\boldsymbol{\eta}}_q-l_q''(\tilde{\boldsymbol{\eta}}_q)^{-1}l_q'(\tilde{\boldsymbol{\eta}}_q)$
  repeat
    For $g=1,2,\ldots,P$: Update $\tilde{\beta}_q^g=\hat{\beta}_q^g$ using (15)
  until: Convergence of $\hat{\boldsymbol{\beta}}_{\boldsymbol{q}}$
  Update $\tilde{\boldsymbol{\beta}}_q=\hat{\boldsymbol{\beta}}_q$
end
Update $\tilde{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}$

3.3 Active set updates

To improve the computational speed of the '**L1mstate**' package, we have constructed an active set that takes advantage of the sparsity of β. As shown in the Algorithm 2, we only need to update the non-zero coefficients $\hat{\beta}_q^g$ in $\boldsymbol{A}$ after a complete cycle has run through all the risk factors, i.e., when $\tilde{\beta}=0$, $\hat{\beta}_q^g$ will stay zero if $\left|-\frac{1}{N_q}l_q'(0)X_{q,g}\right|<\lambda$; otherwise, $\hat{\beta}_q^g$ will be updated and stored in the active set if $\left|-\frac{1}{N_q}l_q'(0)X_{q,g}\right|\ge\lambda$. Therefore, the number of updates is reduced significantly and the convergence of the algorithm is increased. The algorithm will stop if another complete cycle does not change this set. Note that the active set A can only become larger after each update, so the algorithm will always stop after a finite number of updates (See Meier et al. 2007 [34] for more details of the convergence property.)

3.4 Pathwise solution

The above procedure is just for one fixed value of λ. However, in general, it is of interest to be able to compute the optimal solution for a range of values of λ. Thus, we aim to compute the regularization path (denoted as $\hat{\beta}(\lambda)$) where $\lambda\in[0,\infty]$. It can be shown that $\hat{\beta}(\lambda)$ turns out to be a piecewise linear, continuous function of λ [32]. In other words, we only need to compute the solutions on the change points in this path, denoted $\lambda_{max}\ge\lambda_1\ge\cdots\ge\lambda_{min}\ge 0$. We can start with λ_{max} that is any value sufficiently large for which the entire coefficients $\widehat{\beta}=0$. From Eq. (15), notice that when $\widetilde{\beta}=0$, $\hat{\beta}^g$ will stay zero if $\left|-\frac{1}{N_q}l_q'(0)X_{q,g}\right|<\lambda$. Hence, we can set

$$\lambda_{max}=\max\left|-\frac{1}{N_q}l_q'(\mathbf{0})\boldsymbol{X}_{q,g}\right|,\text{for } q=1,2,\ldots,Q;\ g=1,2,\ldots,P.$$

Following the suggestions made in Simon et al. 2011 [41], we can ignore solutions for that are close to 0 and set $\lambda_{min}=\epsilon\lambda_{max}$, then, compute the solutions over $m+1$ values defined as $\lambda_i=\lambda_{max}\left(\frac{\lambda_{min}}{\lambda_{max}}\right)^{\frac{i}{m}}$, for $i=0,1,\ldots,m$ and $\begin{cases}0.01 \text{ if } N<P\\ 0.0001 \text{ if } N\ge P\end{cases}$. In doing this, the algorithm usually converges well because we could use the preceding solution (i.e., for λ_i) as the initial values to obtain the solution for λ_{i-1}.

3.5 Selection of the tuning parameters

With a path of solutions, we need to select an optimal one. The natural choice is cross-validation. However, the partial likelihood of multi-state model is not as well defined as the Gaussian log likelihood on the left-out sample using the traditional cross-validation, which leads to poor results. To tackle it, we used the cross-validation method as described in Verweij *et al.* 1993 [44], proposed for Cox regression model, in which data are split into k parts, use $(k-1)$ parts to train the model, and then, validate the learned model on the whole data. The cross-validated log-partial likelihood for a given part i and λ is

$$\widehat{\mathrm{CV}}_i(\lambda) = l\left(\hat{\boldsymbol{\beta}}_{-i}\right) - l_{i-1}\left(\hat{\boldsymbol{\beta}}_{-i}\right),$$

which can be used as the goodness-of-fit estimate of the solution. Here, $\hat{\beta}$ and l_{-i} are the optimal coefficients and its corresponding log-partial likelihood for data excluding part i. The total goodness-of-fit, $\widehat{CV}(\lambda)$ is the sum of all $\widehat{CV}_i(\lambda)$. We find the optimal λ

$$\hat{\lambda}_{cvl} = \underset{\lambda}{\operatorname{argmax}}\ \widehat{\mathrm{CV}}(\lambda)$$

However, this method alone sometimes produces high true positive rates (TPR) and high false positive rates (FPR). One example of this high positive rates is overfitting. To reduce FPR without large reduction of TPR, we use the penalized method proposed in Ternes *et al.* 2016 [42]. Let p_λ be the number of non-zero coefficients in the model for a given λ, we can find the optimal λ that maximizes

$$\widehat{\mathrm{CV}}(\lambda) - \frac{\widehat{\mathrm{CV}}(\hat{\lambda}_{cvl}) - \widehat{\mathrm{CV}}(\lambda_{max})}{p_{\hat{\lambda}_{cvl}}} * p_\lambda, \text{ for all } \lambda \in \left[\hat{\lambda}_{cvl}, \lambda_{max}\right]$$

Intuitively, it reduces the sparsity of the model p_λ without decreasing much the goodness-of-fit of the model $\widehat{CV}(\lambda)$.

3.6 Estimation of the cumulative hazard rates and the transition probabilities

In the previous section, we have modeled and estimated the effects of the risk factors upon the transition intensities. To further assess the effects of the risk factors on disease progression; in particular, the effects of the risk factors on the cumulative hazard rates and the transition probabilities, we will present how to estimate the transition-specific hazard rates and the transition probabilities in the following.

Given the estimated regression coefficients, the baseline hazards of transition q, denoted by $\alpha_{q0}(t, \beta_q)$, can be obtained as the Breslow estimators [9]

$$\hat{\alpha}_{q0}(t, \hat{\boldsymbol{\beta}}_q) = \frac{dN_q(t)}{S_q^{(0)}(t, \hat{\boldsymbol{\beta}}_q)}$$

where $dN_q(t)$ is the number of events of transition q up to and including time t and

$$S_q^{(0)}(t, \hat{\boldsymbol{\beta}}_q) = \sum_{n=1}^{N_q} \exp\left(\hat{\boldsymbol{\beta}}_q^T \boldsymbol{X}_q^n\right) Y_q^n(t),$$

Recall that, $Y_q^n(t)$ indicates that the n^{th} individual at risk in transition q at time t. Let the risk score for each subject of transition q be $\hat{r}_q^n = \exp\left(\hat{\beta}_q^T X_q^n\right)$, then

$$\hat{\alpha}_{q0}(t, \hat{\boldsymbol{\beta}}_q) = \frac{dN_q(t)}{\sum_{n=1}^{N_q} \hat{r}_q^n Y_q^n(t)},$$

The corresponding estimators of the cumulative baseline hazard $\hat{\Lambda}_{q0}(t, \hat{\beta}_q) = \int_0^t \hat{\alpha}_{q0}(u, \hat{\beta}_q) du$, is computed as

$$\hat{\Lambda}_{q0}(t, \hat{\boldsymbol{\beta}}_q) = \sum_{u \le t} \frac{dN_q(u)}{\sum_{n=1}^{N_q} \hat{r}_q^n Y_q^n(u)},$$

The cumulative hazard rates of transition q, denoted by $\hat{\boldsymbol{\Lambda}}_q(t, \hat{\beta})$ which is also known as the Nelson-Aalen estimators, is

$$\hat{\boldsymbol{\Lambda}}_q(t, \hat{\boldsymbol{\beta}}) = \hat{\Lambda}_{q0}(t, \hat{\boldsymbol{\beta}}_q) \exp\left(\hat{\boldsymbol{\beta}}_q^T \boldsymbol{X}_q\right),$$

Given the cumulative transition hazards, using the basic tool − a product integral allows us to estimate the transition probability matrix $P(s,t) = \mathrm{P}_{hj}(s,t)$ as

$$\mathrm{P}(s,t) = \prod_{u \in (s,t]} \left(\mathrm{I} + \boldsymbol{\Delta}\hat{\boldsymbol{\Lambda}}(u)\right)$$

where $\prod_{u \in (s,t]}$. is a product-integral and $(s,t]$ denotes the time interval. It is the Aalen-Johansen estimator [2].

3.7 Computational complexity analysis

We now discuss the complexity of the algorithms when using different frameworks (L1MSTATE, L1Cox, L1-StratifiedCox) for variable selection. They all solve the optimization problems by the coordinate descent algorithms to optimize the objective function with respect to one variable at a time while all the others are fixed. In other words, they process the same procedure: precompute the first-order derivatives and the diagonal entries of the second-order derivatives of a design matrix; at each iteration update P_a – the number of nonzero elements in the active set. The computational complexity depends on the number of subjects N, the number of risk factors P and the number of transitions Q. More specifically, consider L1MSTATE and L1Cox, for each transition, they need $O(N_q^2)$ operations to compute the derivatives where N_q is the number of subjects for transition q (recall that $N = \sum_{q=1}^{Q} N_q$) and each update needs $O(P)$ operations. Therefore, their complexity is $O\left(\sum_{q=1}^{Q}\left(N_q^2 + P_q^a P\right)\right)$ where P_q^a denotes the number of nonzero elements of transition q. For L1-StratifiedCox, it needs to create transition-specific risk factors from the baseline risk factors as described in [13]: each risk factor X is split into as many risk factors X_q as there are transitions in the model, for transition q, $X_q = X$; while for all other transitions $X_q = 0$. It means that the number of risk factors now is PQ. In addition, it needs $O(N^2)$ operations to compute the derivatives. Therefore, its complexity is $O\left(N^2 + PQ\sum_{q=1}^{Q} P_q^a\right)$. Of course, the required runtime for the entire solution path also depends on the number of iterations, which in turn depends on the data and λ values. In general, the dominant factor influencing the number of iterations is the number of nonzero elements at the specific λ value since the nonactive elements that remain fixed at zero need no iteration. In the next section, we compare their computational complexity empirically in Table 9 with the runtime of three L1-regularized models using the same maximum number of iterations 10^5 for all models.

4 Simulation Studies

In this section, we will numerically compare the performance of the L1-regularized multi-state model (L1MSTATE) with existing regularized multi-state models including the L1-regularized cause-specific Cox proportional hazards model (L1Cox) that is commonly used in survival analysis without multistate structure knowledge, and the L1-regularized stratified Cox proportional hazards model (L1-StratifiedCox) in term of variable selection using simulated data. The L1-regularized estimation of the fused-lasso multi-state model approach [38] was not included in our comparison due to very huge computation cost (see Discussion section for more details.)

To compare the performance of the four models in terms of identification of the significant risk factors, we calculated three performance metrics, including the true positive rate (TPR), false positive rate (FPR), and area under the ROC curve (AUC).

4.1 Setup

Following the data structure outlined in Section 2.4, we generate trajectories of N individuals that include their transitions among states, the times of the transitions, and the values of risk factors. First, the values of the risk factors of each individual are generated by randomly sampling from a P-dimensional multivariate normal distribution with mean vector as zero and the correlation matrix $\mathbf{C}$ as an autoregressive matrix where $\mathbf{C}_{ij} = \rho^{|i-j|}$ and $0 \leq \rho \leq 1$. The reason to use an autoregressive correlation matrix is that we could flexibly tune the correlations of the variables by setting the value of ρ, i.e., $\rho = 0$ means no correlation among the variables, while $\rho = 1$ means that the risk factors are perfectly correlated as duplicates of each other. Second, the transitions among states and their timing are generated as follows. Recall that we have assumed that the transition intensities between two states follow the proportional hazards Cox model Eq. (1). By setting up values for β we can obtain the transition intensity distribution from Eq. (1) to randomly sample the transition intensity values. After that, the observed times of the transition events between two states are generated using the exponential distribution with its rate parameter set to be the transition intensity between these two states. In here, we consider the illness-death model that includes three states: **healthy**, **illness**, and **death**. Its transition structure is depicted in Figure 2.

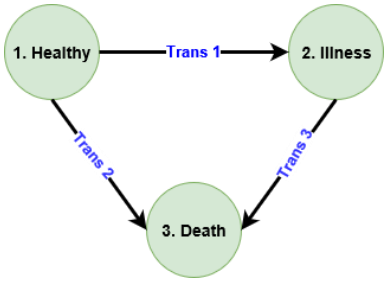


Fig. 2: The illness-death model.

Assume that all individuals start at the healthy state in the beginning of the observation period. The censoring status values are generated as follows. Since the observation time is the exact transition time, there is no illness censoring time or the censoring indicator of transition to illness state is 1 for all N individuals. The death censoring times are generated from the exponential distribution, the censoring indicator of transitions to death state is 0 if the death time is larger than the death censoring time, and 1 otherwise. The strength of effect of risk factor is based on the real absolute value of its corresponding coefficient. Next, we

first consider the small datasets in which the number of subjects and the number of risk factors are relatively small. We further test different methods on the datasets with a large number of subjects and risk factors.

4.2 Small datasets

4.2.1 Settings

In these setting, we include the un-regularized multi-state model (MSTATE) to investigate the pros and cons of the un-regularized methods comparing with the regularized methods.

Set the number of risk factors $P = 9$ and different values of sample size, i.e., $N \in \{100, 250, 450\}$. We consider four scenarios: the first three scenarios include the effects of risk factors belong the same type (large, medium, or small), and the last scenario includes all three types of the effects of risk factors.

- First scenario: small effects

$$\boldsymbol{\beta} = \begin{bmatrix} 0.15 & 0.15 & 0.15 & 0 & 0 & 0.15 & 0.15 & 0 & 0 \\ 0.15 & 0.15 & 0 & 0 & 0 & 0 & 0.15 & 0 & 0 \\ 0 & 0.15 & 0.15 & 0 & 0 & 0.15 & 0.15 & 0 & 0.15 \end{bmatrix}$$

- Second scenario: medium effects

$$\boldsymbol{\beta} = \begin{bmatrix} -0.35 & -0.35 & -0.35 & 0 & 0 & -0.35 & -0.35 & 0 & 0 \\ -0.35 & -0.35 & 0 & 0 & 0 & 0 & -0.35 & 0 & 0 \\ 0 & -0.35 & -0.35 & 0 & 0 & -0.35 & -0.35 & 0 & -0.35 \end{bmatrix}$$

- Third scenario: large effects

$$\boldsymbol{\beta} = \begin{bmatrix} -0.65 & -0.65 & -0.65 & 0 & 0 & -0.65 & -0.65 & 0 & 0 \\ -0.65 & -0.65 & 0 & 0 & 0 & 0 & -0.65 & 0 & 0 \\ 0 & -0.65 & -0.65 & 0 & 0 & -0.65 & -0.65 & 0 & -0.65 \end{bmatrix}$$

- Fourth scenario: mixed effects

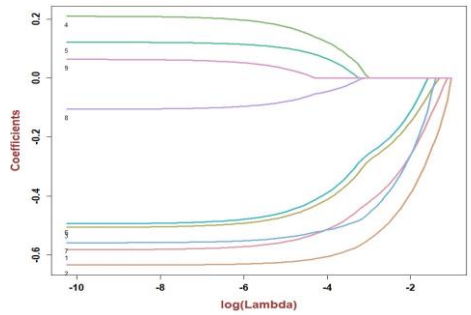
$$\boldsymbol{\beta} = \begin{bmatrix} 0.15 & -0.35 & -0.35 & 0 & 0 & -0.35 & -0.35 & 0 & 0 \\ 0 & 0.15 & -0.65 & 0 & 0 & 0 & -0.65 & 0 & 0 \\ 0 & -0.65 & -0.65 & 0 & 0 & -0.35 & -0.65 & 0 & 0.15 \end{bmatrix}$$

We evaluate different levels of correlation between the risk factors by setting $\rho = 0, 0.2, 0.5$. The censoring percentage is 30%.

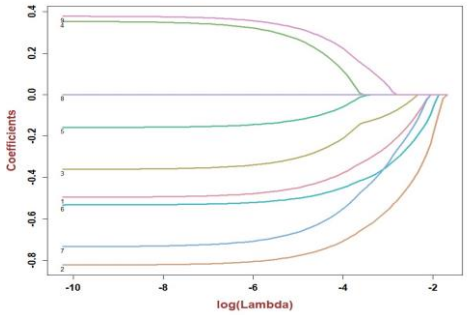
4.2.2. Results

To compute TPRs and FPRs for the disease progression from the healthy state to the death state for our L1MSTATE, we created a path of 100 values of λ, applied 10-fold for two different cross-validation methods described above in Section 3.5 to select the optimal λ for variable selection. We can view the estimated coefficients from our L1MSTATE model fit, and the cross-validation log-partial likelihood against the log of λ values, and also how to use two different cross-validation methods to select λ. Figure 3 shows the results of the large effects setting in which $N = 250$, and $\rho = 0.5$. For L1Cox and L1-StratifiedCox, we used '**glmnet**' package [41] with its default setting to fit Cox proportional hazards models: 100 values of λ and 10-fold cross-validation, which is the same as the first cross-validation method used in our model, to select the optimal solution. More specifically, for L1Cox, we applied for each transition using transition-specific datasets, then used the results of three transitions to compute the TPRs and FPRs; for L1-StratifiedCox, we applied directly to the long-format data. For MSTATE, we used R package '**mstate**' [13] to fit model and the statistical hypothesis test (p-value) with the 0.05 significance level to evaluate the significance of candidate risk factors based on Wald tests on each variable for variable selection instead of using some methods such as backward or forward selection. The results across 100 replications for these models in different scenarios are summarized in Tables 2, 3, 4, and 5.

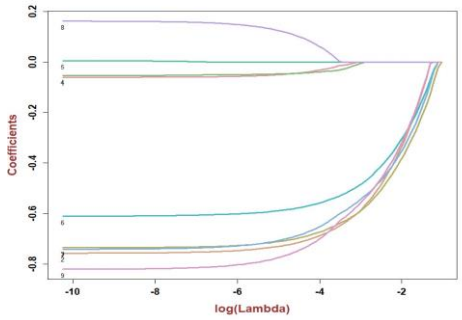
The results from Tables 2, 3, 4, and 5 show that TPR and FPR values of pL1MSTATE are always lower than L1MSTATE. It means that the penalized cross-validation method is more conservative than the first cross-validation method. On the one hand, comparing L1MSTATE and MSTATE results, MSTATE always gives lower TPRs and FPRs than L1MSTATE. In other words, applying the statistical hypothesis test with the 0.05 significance level to MSTATE produces more sparse models than applying the first cross-validation method to L1MSTATE. On the other hand, comparing pL1MSTATE and MSTATE results shows that when $N = 100$ pL1MSTATE often gives lower both TPRs and FPRs than MSTATE; when N = 250 and N = 450 in small setting, pL1MSTATE gives better results than MSTATE; in other settings, pL1MSTATE starts giving lower both TPRs and FPRs than MSTATE, and MSTATE gives better results in large effects setting. Note that when ρ increases - risk factors become highly correlated, MSTATE results become worse while L1MSTATE and pL1MSTATE results often become better. Consider three regularized models L1MSTATE, L1Cox, and L1-StratifiedCox using the same cross-validation method, from Tables 2, 3, 4 and 5, it can be seen that L1MSTATE is always better than L1-StraitifiedCox. Compare L1MSTATE and L1Cox: when $N = 100$, L1MSTATE is more conservative than L1Cox since it gives both the smaller TPRs and FPRs; when N increases, L1MSTATE gives the better results with the higher TPRs and the lower FPRs; when $N = 450$ in the large effects case, three regularized models perform the same.



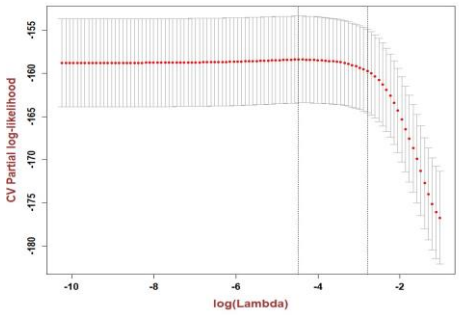
(a) Transition 1 → 2



(b) Transition 1 → 3



(c) Transition 2 → 3



(d) The cross-validation curve (red dotted line), and its standard deviation

Fig. 3: Plots of the coefficient paths for three transitions of our L1MSTATE model fit and the cross-validation log-partial likelihood against the log of λ values along our path. In the first three plots, each curve corresponds to a risk factor and is annotated by index of this risk factor. In the final plot, each dot represents the log of λ values along the path, and error bars give a confidence interval for the cross-validation log-partial likelihood. The left vertical bar indicates the maximum cross-validation partial-log-likelihood while the right one shows the penalized cross-validation log-partial likelihood.

Table 2: Model selection results of Example I for the small effects scenario.

N	ρ	MSTATE		pL1MSTATE		L1MSTATE		L1Cox		L1-StratifiedCox	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
100	0	0.17	0.08	0.14	0.05	0.20	0.09	0.31	0.20	0.20	0.12
	0.2	0.18	0.08	0.20	0.07	0.26	0.11	0.39	0.25	0.26	0.14
	0.5	0.14	0.08	0.25	0.10	0.36	0.18	0.45	0.30	0.39	0.22
250	0	0.32	0.07	0.30	0.08	0.51	0.27	0.61	0.38	0.54	0.30
	0.2	0.28	0.06	0.42	0.12	0.62	0.30	0.67	0.44	0.62	0.33
	0.5	0.22	0.06	0.45	0.13	0.70	0.33	0.73	0.45	0.70	0.39
450	0	0.47	0.08	0.52	0.11	0.84	0.45	0.83	0.56	0.83	0.49
	0.2	0.47	0.08	0.56	0.10	0.85	0.42	0.85	0.53	0.86	0.47
	0.5	0.37	0.06	0.59	0.13	0.86	0.43	0.87	0.50	0.83	0.44

pL1MSTATE, L1-regularized multi-state model using the penalized cross validation method; L1MSTATE, L1-regularized multi-state model using the first cross-validation method; MSTATE, multi-state model; L1Cox, L1regularized cause-specific Cox model using the first cross validation method; L1-StratifiedCox, L1-regularized stratified Cox model using the first cross-validation method; TPR, true positive rate; FPR, false positive rate.

The TPRs and FPRs shown in these above tables depend on the selected methods including the cross-validation methods, and the significance level of $p-$value. We want to evaluate further the variable selection performance of these models using the area under a curve (AUC) values that are also variable selection metrics and do not depend on the selected methods. We use the same settings as above with different values of sample size, i.e., $N \in \{50, 75, \ldots, 500\}$. We first calculate the TPRs and FPRs, then compute the AUC values by using the method described in Fawcett *et al.* 2006 [16]. Intuitively, the TPR and FPR pairs were calculated to construct ROC curves, then the area under a ROC curve (AUC) was computed. More specifically,

Table 3: Model selection results of Example I for the medium effects scenario.

N	ρ	MSTATE		pL1MSTATE		L1MSTATE		L1Cox		L1-StratifiedCox	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
100	0	0.50	0.08	0.50	0.13	0.70	0.30	0.78	0.50	0.70	0.38
	0.2	0.43	0.07	0.50	0.12	0.76	0.33	0.80	0.46	0.76	0.40
	0.5	0.32	0.08	0.56	0.14	0.80	0.38	0.83	0.50	0.80	0.43
250	0	0.84	0.10	0.81	0.15	0.98	0.60	0.98	0.65	0.98	0.64
	0.2	0.83	0.08	0.82	0.13	0.98	0.58	0.99	0.63	0.98	0.59
	0.5	0.70	0.07	0.77	0.13	0.97	0.52	0.97	0.60	0.97	0.53
450	0	0.96	0.14	0.88	0.13	1.00	0.69	1.00	0.70	1.00	0.69
	0.2	0.97	0.12	0.91	0.13	1.00	0.65	1.00	0.68	1.00	0.66
	0.5	0.88	0.10	0.87	0.14	1.00	0.59	1.00	0.62	0.99	0.61

Table 4: Model selection results of Example I for the large effects scenario.

N	ρ	MSTATE		pL1MSTATE		L1MSTATE		L1Cox		L1-StratifiedCox	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
100	0	0.84	0.12	0.80	0.17	0.97	0.53	0.97	0.62	0.97	0.60
	0.2	0.81	0.12	0.78	0.13	0.97	0.50	0.98	0.59	0.97	0.57
	0.5	0.67	0.10	0.82	0.20	0.97	0.50	0.98	0.58	0.96	0.54
250	0	0.99	0.16	0.92	0.14	1.00	0.70	1.00	0.70	1.00	0.70
	0.2	0.98	0.13	0.94	0.15	1.00	0.66	1.00	0.68	1.00	0.66
	0.5	0.96	0.11	0.93	0.18	1.00	0.59	1.00	0.61	1.00	0.62
450	0	1.00	0.19	0.97	0.15	1.00	0.73	1.00	0.73	1.00	0.73
	0.2	1.00	0.19	0.97	0.12	1.00	0.70	1.00	0.70	1.00	0.70
	0.5	0.99	0.15	0.96	0.18	1.00	0.65	1.00	0.65	1.00	0.66

Table 5: Model selection results of Example I for the mixed effects scenario.

N	ρ	MSTATE		pL1MSTATE		L1MSTATE		L1Cox		L1-StratifiedCox	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
100	0	0.47	0.08	0.40	0.07	0.67	0.29	0.70	0.37	0.67	0.36
	0.2	0.47	0.06	0.40	0.06	0.71	0.34	0.74	0.37	0.71	0.39
	0.5	0.36	0.07	0.45	0.11	0.71	0.34	0.75	0.42	0.70	0.40
250	0	0.68	0.08	0.56	0.06	0.89	0.51	0.90	0.55	0.88	0.56
	0.2	0.67	0.09	0.53	0.05	0.88	0.50	0.88	0.58	0.86	0.51
	0.5	0.59	0.07	0.56	0.11	0.86	0.49	0.87	0.53	0.84	0.51
450	0	0.77	0.10	0.62	0.04	0.95	0.61	0.97	0.64	0.95	0.63
	0.2	0.75	0.10	0.60	0.04	0.96	0.61	0.96	0.66	0.95	0.63
	0.5	0.71	0.08	0.58	0.07	0.92	0.56	0.91	0.58	0.89	0.56

pL1MSTATE, L1-regularized multi-state model using the penalized cross validation method; L1MSTATE, L1-regularized multi-state model using the first cross-validation method; MSTATE, multi-state model; L1Cox, L1regularized cause-specific Cox model using the first cross validation method; L1-StratifiedCox, L1-regularized stratified Cox model using the first cross-validation method; TPR, true positive rate; FPR, false positive rate.

in three regularized models L1MSTATE, L1Cox, and L1-StratifiedCox, it is straightforward to calculate 100 pairs of TPRs and FPRs corresponding to 100 λ values along the path. In MSTATE, the threshold path was constructed, and it included only the corresponding $p-$values of estimated coefficients. Then, for each threshold, the risk factors that have smaller $p-$values than the threshold were selected, and the corresponding TPR and FPR pairs were computed. The results of AUC values of these models in twelve settings for different datasets over 100 replications are shown in Figure 4.

First, we compare the performances of L1MSTATE and MSTATE. From Figure 4, in small effects setting, L1MSTATE gives comparable performance with MSTATE when there is no correlation among risk factors ($\rho = 0$), and better performance than MSTATE when the correlation ρ becomes higher. Other settings show the same pattern: when sample size is small, MSTATE performs worse than L1MSTATE; when sample size increases, MSTATE's performance gradually catches up, and even becomes better than L1MSTATE's performance. Notice that when the correlation among risk factors ρ increases, MSTATE needs more samples to be able to catch up L1MSTATE's performance, and when the effects become stronger, MSTATE needs less samples to perform comparably with L1MSTATE.

Second, we compare the performance of three regularized models L1MSTATE, L1Cox, and L1-StratifiedCox. In the first three settings L1MSTATE always gives the best performance. In the last setting L1MSTATE gives slightly worse performance than L1Cox when $\rho = 0$, and comparable when ρ increases; L1MSTATE also gives better performance than L1-StratifiedCox. Two models L1Cox and L1-StratifiedCox perform differently: they perform comparably in small effects setting; L1-StratifiedCox performs better L1Cox in medium and large effects settings; L1Cox performs better L1-StratifiedCox in mixed effects setting. L1MSTATE performs better than L1Cox can most likely be explained by the benefit of incorporating the prior knowledge about the disease progression model: in L1MSTATE, we incorporated information about multi-state model of disease progression into data process when converting the original data to long-format data; L1Cox, by contrast, applied L1-regularized Cox proportional hazards model for each transition-specific dataset separately.

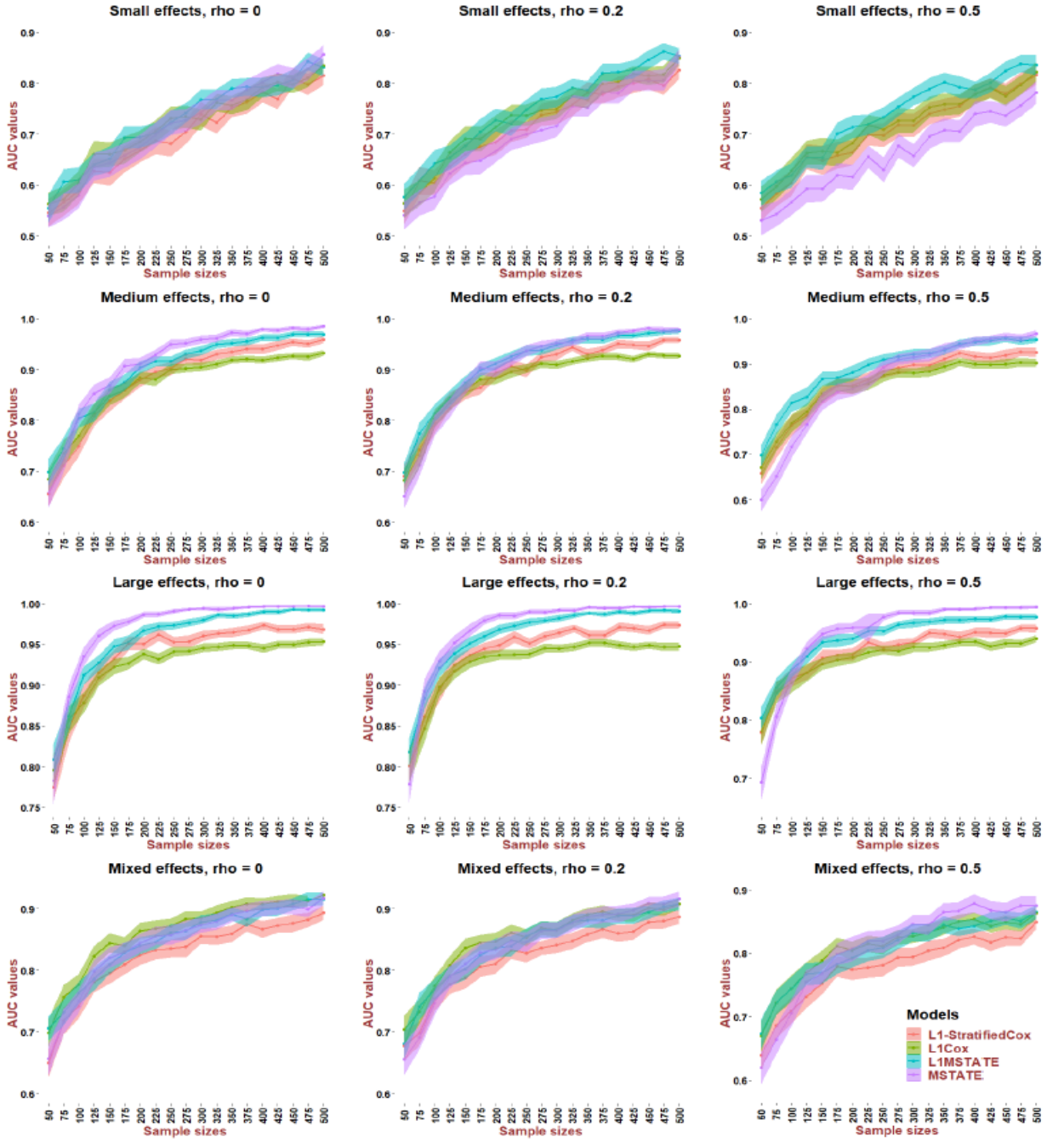


Fig. 4: AUC values of Example I for different sample sizes in different settings over 100 replications.

Table 6: Model selection results of large-scale datasets for the small effects scenario.

		L1MSTATE			L1Cox			L1-StratifiedCox		
N	ρ	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC
3000	0.5	0.75	0.38	0.76	0.78	0.43	0.74	0.73	0.39	0.72
	0.2	0.76	0.42	0.75	0.73	0.51	0.74	0.78	0.45	0.72
	0	0.74	0.43	0.73	0.72	0.52	0.73	0.78	0.48	0.71
6000	0.5	0.87	0.51	0.81	0.88	0.55	0.79	0.86	0.50	0.77
	0.2	0.94	0.64	0.81	0.96	0.68	0.80	0.94	0.63	0.80
	0	0.95	0.69	0.80	0.96	0.73	0.78	0.95	0.69	0.79
9000	0.5	0.92	0.57	0.83	0.92	0.62	0.81	0.92	0.52	0.80
	0.2	0.98	0.72	0.84	0.98	0.74	0.82	0.98	0.67	0.83
	0	0.99	0.76	0.82	0.99	0.78	0.81	0.99	0.73	0.82

Table 7: Model selection results of large-scale datasets for the medium effects scenario.

		L1MSTATE			L1Cox			L1-StratifiedCox		
N	ρ	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC
3000	0.5	0.89	0.69	0.78	0.81	0.51	0.77	0.77	0.39	0.73
	0.2	0.91	0.64	0.79	0.49	0.60	0.78	0.88	0.53	0.77
	0	0.88	0.58	0.78	0.91	0.64	0.78	0.92	0.60	0.76
6000	0.5	0.95	0.77	0.82	0.89	0.66	0.81	0.90	0.50	0.78
	0.2	0.98	0.67	0.84	0.97	0.77	0.83	0.98	0.67	0.82
	0	0.99	0.74	0.83	0.99	0.81	0.82	0.99	0.73	0.82
9000	0.5	0.96	0.60	0.86	0.96	0.72	0.84	0.96	0.55	0.82
	0.2	0.99	0.72	0.86	0.99	0.82	0.85	0.99	0.70	0.84
	0	1	0.75	0.85	1	0.85	0.84	1	0.76	0.84

Table 8: Model selection results of large-scale datasets for the large effects scenario.

		L1MSTATE			L1Cox			L1-StratifiedCox		
N	ρ	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC
3000	0.5	0.83	0.47	0.77	0.82	0.57	0.77	0.78	0.41	0.73
	0.2	0.96	0.77	0.80	0.89	0.66	0.79	0.90	0.54	0.78
	0	0.99	0.86	0.80	0.93	0.71	0.79	0.94	0.62	0.77
6000	0.5	0.94	0.54	0.84	0.90	0.72	0.81	0.90	0.50	0.80
	0.2	0.98	0.66	0.84	0.97	0.82	0.83	0.98	0.66	0.82
	0	1	0.76	0.85	1	0.86	0.84	1	0.73	0.83
9000	0.5	0.96	0.48	0.83	0.96	0.79	0.84	0.95	0.55	0.82
	0.2	0.99	0.67	0.87	0.99	0.87	0.85	0.99	0.69	0.84
	0	1	0.74	0.87	1	0.90	0.85	1	0.75	0.84

L1MSTATE, L1-regularized multi-state model using the first cross-validation method; L1Cox, L1regularized cause-specific Cox model using the first cross validation method; L1-StratifiedCox, L1-regularized stratified Cox model using the first cross-validation method; TPR, true positive rate; FPR, false positive rate; AUC, area under a curve.

L1MSTATE performs better than L1-StratifiedCox even though both L1MSTATE and L1-StratifiedCox models use long-format data. The reason is that L1MSTATE assigned different weights to each transition while L1-StratifiedCox did not. Intuitively, L1MSTATE put higher penalties on rare transitions than common transitions.

In summary, the L1-regularized multi-state model (L1MSTATE) is the best one among the regularized models in terms of variable selection. L1MSTATE performs better at variable selection than the un-regularized multi-state model (MSTATE) when sample sizes are small or the effects are small, and MSTATE performs better than L1MSTATE when sample sizes are large or the effects are strong.

4.3 Large-scale datasets

In this setting, we only compare the performances of three L1-regularized models without including the un-regularized multi-state model (MSTATE). We set the number of risk factors $P = 300$ and the number of nonzero ones to be 100 per each transition. Different sample sizes, i.e., $N \in \{3000, 6000, 9000\}$, are simulated. The results of three L1-regularized models are shown in Tables 6, 7 and 8. They are consistent with the results of small datasets, which suggests that L1MSTATE is better than L1Cox and L1-StratifiedCox in terms of accurate variable selection.

4.4 Empirical runtime comparison

We further compare the runtime of three L1-regularized multi-state models on all the simulated datasets. As shown in Table 9, our L1MSTATE is the most computational efficient as we expected based on our previous computational complexity analysis.

Table 9: Running time of three L1-regularized models. The mean time over different datasets (100 for small datasets and 10 for big datasets) required to fit the entire solution path over a grid of 100 λ values is reported in seconds.

N	ρ	L1MSTATE			L1Cox			L1-StratifiedCox		
		Small	Medium	Large	Small	Medium	Large	Small	Medium	Large
100	0	0.01	0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.03
	0.2	0.01	0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.04
	0.5	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.04	0.04
250	0	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.05	0.05
	0.2	0.02	0.02	0.02	0.03	0.03	0.04	0.04	0.05	0.05
	0.5	0.02	0.02	0.02	0.03	0.03	0.04	0.05	0.05	0.06
450	0	0.03	0.03	0.03	0.04	0.05	0.05	0.07	0.07	0.08
	0.2	0.03	0.03	0.03	0.04	0.05	0.06	0.07	0.07	0.08
	0.5	0.03	0.03	0.04	0.05	0.05	0.07	0.07	0.09	0.10
3000	0	2.47	3.29	5.12	3.93	4.72	6.49	11.43	11.79	12.51
	0.2	2.53	4.20	4.27	3.94	5.14	6.43	11.34	11.92	11.53
	0.5	2.96	4.42	3.30	4.46	5.97	8.06	12.13	12.49	12.65
6000	0	4.78	7.93	6.71	7.24	9.15	10.27	21.07	25.07	21.18
	0.2	6.22	7.97	5.56	7.90	9.57	10.56	24.87	24.71	20.64
	0.5	5.28	9.95	5.32	8.19	11.31	13.58	22.93	26.84	22.92
9000	0	8.48	12.34	12.71	10.91	13.15	17.09	33.16	37.65	38.77
	0.2	8.28	11.48	10.82	11.13	13.97	17.23	33.63	33.86	37.37
	0.5	9.06	9.10	9.39	12.15	15.64	21.01	35.72	35.83	37.62

L1MSTATE, L1-regularized multi-state model using the first cross-validation method; L1Cox, L1regularized cause-specific Cox model using the first cross validation method; L1-StratifiedCox, L1-regularized stratified Cox model using the first cross-validation method

5 Europe Blood and Marrow Transplantation (EBMT) data

In this section, we will compare the performance of L1-regularized multi-state model (L1MSTATE) with un-regularized multi-state model (MSTATE) in terms of the predictions of the transition probabilities, and demonstrate how to use our '**L1mstate**' package to further assess the effects of risk factors upon the disease progression using the Europe Blood and Marrow Transplantation (EBMT) dataset that has been described and analyzed in deWreede *et al.* 2011 [13].

The model for the leukemia patients after bone marrow transplantation (so-called EBMT model) is shown in Figure 5. The EBMT model includes six states and twelve possible transitions. These states are transplant (Tx) state, recovery (Rec) state, adverse event (AE) state, combination of adverse event and recovery state (Rec+AE), relapse (Rel) state, and death, respectively. The numeric coding $1, 2, \ldots, 12$ represent twelve possible transitions. This dataset includes 2279 patients and the observed transitions are summarized in Table 10.

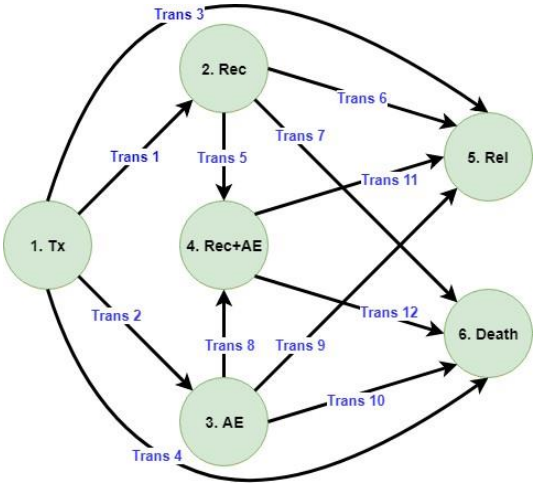


Fig. 5: The EBMT model.

Table 10: The frequencies and proportions of the number of observed transitions of study population. The numbers in parentheses are proportions.

	Tx	Rec	AE	Rec+AE	Rel	Death	No event	Total
Tx	0 (0)	785 (0.34)	907 (0.40)	0 (0)	95 (0.04)	160 (0.07)	332 (0.15)	2279
Rec	0 (0)	0 (0)	0 (0)	227 (0.29)	112 (0.14)	39 (0.05)	407 (0.52)	785
AE	0 (0)	0 (0)	0 (0)	433 (0.48)	56 (0.06)	197 (0.22)	221 (0.24)	907
Rec+AE	0 (0)	0 (0)	0 (0)	0 (0)	107 (0.16)	137 (0.21)	416 (0.63)	660

The six risk factors are donor-recipient match, prophylaxis, year of transplant, and age of transplant in years. All of them are categorical risk factors. As in this paper we focus on numeric risk factors, we convert them to numeric by using dummy coding as follow

x_1: donor-recipient match (1 refers to yes and 0 refers to no)

x_2: prophylaxis (1 refers to yes and 0 refers to no)

x_3: year of transplant (1 refers to 1990-1994 and 0 refers to 1985-1989 or 1995-1998)

x_4: year of transplant (1 refers to 1995-1998 and 0 refers to 1985-1989 or 1990-1994)

x_5: age of transplant (1 refers to 20-40 and 0 refers to < 20 or > 40)

x_6: age of transplant (1 refers to > 40 and 0 refers to < 20 or 20-40)

There are 12 allowable transitions in the model and six time-fixed risk factors for all transitions, resulting in the total number of coefficients as large as 72. For L1MSTATE, we used the regularization path of 100 values of λ, and applied 10-fold for both the first cross-validation method and the penalized cross-validation method to tune the penalty parameter λ. For MSTATE model, we used p-values to select the significant risk factors (highlighted as bold in Table 11). The results from two models in Table 11 are consistent with the results in simulation studies: the penalized cross-validation method is more conservative than the first cross-validation method since it chooses more sparse multi-state models.

5.1 Comparison of the models

We compared the performance of L1MSTATE and MSTATE in terms of the predictions of the transition probabilities. As discussed in the introduction, our aim is to study how L1MSTATE and MSTATE predict the *rare* transitions that have relatively small number of observations and the *common* transitions that have relatively large number of observations. To do it, the transitions from the transplant state were considered, and three example patients A, B, and C (see Table 12) were chosen. The observed transitions from the transplant state of three patients are summarized in Table 13. The summary shows that the transitions from the transplant state to the recovery state and adverse event state have relatively large number of observations while the transition from the transplant state to the relapse state and the death state have relatively small number of observations. In other words, the transitions from the transplant state to the recovery state and adverse event state are the *common* transitions and the transition from the transplant state to the relapse state and the death state are *rare* transitions. In addition, patient A has the largest number of observations (287) that represents the large sample size case and patient C has the smallest number of observations (50) that represents the small sample size case. The same Aalen-Johansen method to predict the transition probabilities were used in both L1MSTATE and MSTATE. The results are shown in Figures 6: the probabilities of transitions from the transplant state at the starting computation times 0 to the ending computation times are estimated and stacked together where the distance between two adjacent curves shows the probability of the state whose name is labeled.

Table 11: Regression coefficients of two models for EBMT dataset.

Methods	Risk factors	Transitions											
		1	2	3	4	5	6	7	8	9	10	11	12
MSTATE	x_1	**-0.167**	-0.111	0.196	-0.003	0.190	**0.426**	0.244	0.126	-0.414	0.008	-0.301	**0.572**
	x_2	**-0.366**	**-0.278**	0.385	-0.056	-0.282	0.268	-0.008	0.125	0.159	0.324	0.012	-0.112
	x_3	**0.401**	0.023	0.442	-0.359	-0.095	-0.210	**-0.836**	**0.528**	-0.311	**-0.644**	-0.024	-0.362
	x_4	**0.521**	-0.114	0.221	**-0.476**	-0.151	0.055	**-0.980**	**0.930**	-0.580	-0.213	-0.390	-0.352
	x_5	0.049	0.123	-0.094	**0.766**	0.292	-0.255	0.150	**-0.393**	0.172	0.238	0.414	**0.760**
	x_6	0.199	0.067	-0.232	**0.934**	**0.470**	-0.101	**1.465**	**-0.328**	0.423	**0.495**	0.256	**1.337**
pL1MSTATE	x_1	-0.040	0	0	0	0	0.100	0	0	0	0	0	0.374
	x_2	-0.291	-0.137	0	0	-0.256	0	0	0	0	0.241	0	0
	x_3	0.117	0	0	0	0	0	0	0.231	0	-0.378	0	0
	x_4	0.250	-0.002	0	0	0	0	0	0.604	0	0	-0.080	0
	x_5	0	0	0	0	0.080	0	0	-0.193	0	0	0.035	0.106
	x_6	0.082	0	0	0	0.178	0	0.460	-0.056	0	0.085	0	0.627
L1MSTATE	x_1	-0.147	-0.093	0.053	0	0.146	0.385	0	0.068	-0.269	0	-0.262	0.521
	x_2	-0.352	-0.253	0.183	0	-0.464	0.199	0	0.071	0.161	0.315	0	-0.062
	x_3	0.353	0.022	0.183	-0.154	-0.068	-0.207	-0.501	0.458	-0.110	-0.602	0	-0.27
	x_4	0.476	-0.091	0	-0.230	-0.050	0.022	-0.593	0.850	-0.267	-0.153	-0.304	-0.241
	x_5	0.007	0.082	0	0.486	0.414	-0.165	0	-0.370	0	0.155	0.305	0.615
	x_6	0.158	0.018	0	0.568	0.557	0	1.154	-0.288	0.147	0.398	0.137	1.196

pL1MSTATE, L1-regularized multi-state model using the penalized cross-validation method; L1MSTATE, L1-regularized multistate model using the first cross-validation method; MSTATE, multi-state model. For MSM method, the significance of risk factors that are at 0.05 levels are shown in bold.

Table 12: Risk factors information of patient A, B, C and D.

Risk factors	Patient A	Patient B	Patient C	Patient D
x_1	0	0	1	0
x_2	0	0	0	0
x_3	1	0	1	0
x_4	0	1	0	1
x_5	1	0	0	1
x_6	0	1	0	0

Table 13: The frequencies and proportions of the number of observed transitions from the transplant state of three patients. The numbers in parentheses are proportions.

		Tx	Rec	AE	Rec+AE	Rel	Death	No event	Total
Patient A	Tx	0 (0)	99 (0.34)	133 (0.46)	0 (0)	9 (0.03)	13 (0.05)	33 (0.11)	287
Patient B	Tx	0 (0)	56 (0.38)	60 (0.41)	0 (0)	5 (0.03)	9 (0.06)	17 (0.12)	147
Patient C	Tx	0 (0)	22 (0.44)	16 (0.32)	0 (0)	3 (0.06)	4 (0.08)	5 (0.10)	50

From Figure 6, it can be seen that the predicted probabilities from the transplant state at the starting time 0 of patient A using MSTATE, pL1MSTATE and L1MSTATE are almost similar but MSTATE and L1MSTATE slightly underestimates the probability of the relapse (Rel) state, and pL1MSTATE slightly underestimates the probability of the adverse event (AE) state comparing with the observed probability. In other words, pL1MSTATE slightly underestimates the probability of the common event while MSTATE and L1MSTATE slightly underestimates the probability of the rare event. The results of patient B clearly shows that MSTATE overestimates the probability of the common event - the recovery (Rec) state, and underestimates the probability of the rare event - the relapse (Rel) state. L1MSTATE also overestimates the probability of the recovery (Rec) state. pL1MSTATE gives the best overall performance. The results of patient C show the same pattern: MSTATE underestimates the probability of the rare event - the death state, and inaccurate prediction of the probability of the relapse state. By contrast, pL1MSTATE and L1MSTATE produces better predictions of these two rare events. The figure also indicates that MSTATE, pL1MSTATE and L1MSTATE overestimate the probability of the common event - the adverse event (AE) state.

In short, the un-regularized multi-state model (MSTATE) tends to underestimate the probabilities of the rare transitions, and overestimate the probabilities of the common transitions. Its performance becomes worse when the sample size decreases. In these cases, our L1-regularized multi-state models produce better predictions.

5.2 Further assessment of the effects of risk factors upon the disease progression

We illustrate how to use the functions of our '**L1mstate**' package to estimate the cumulative hazard rates and the transition probabilities. For illustrative purposes, we continued using patient A example, and chose another patient D (in Table 12) that differs from patient A only in terms of year of transplant since our aim is to assess the effect of year of transplant. The penalized cross-validation method was implemented to select the optimal tuning parameters.

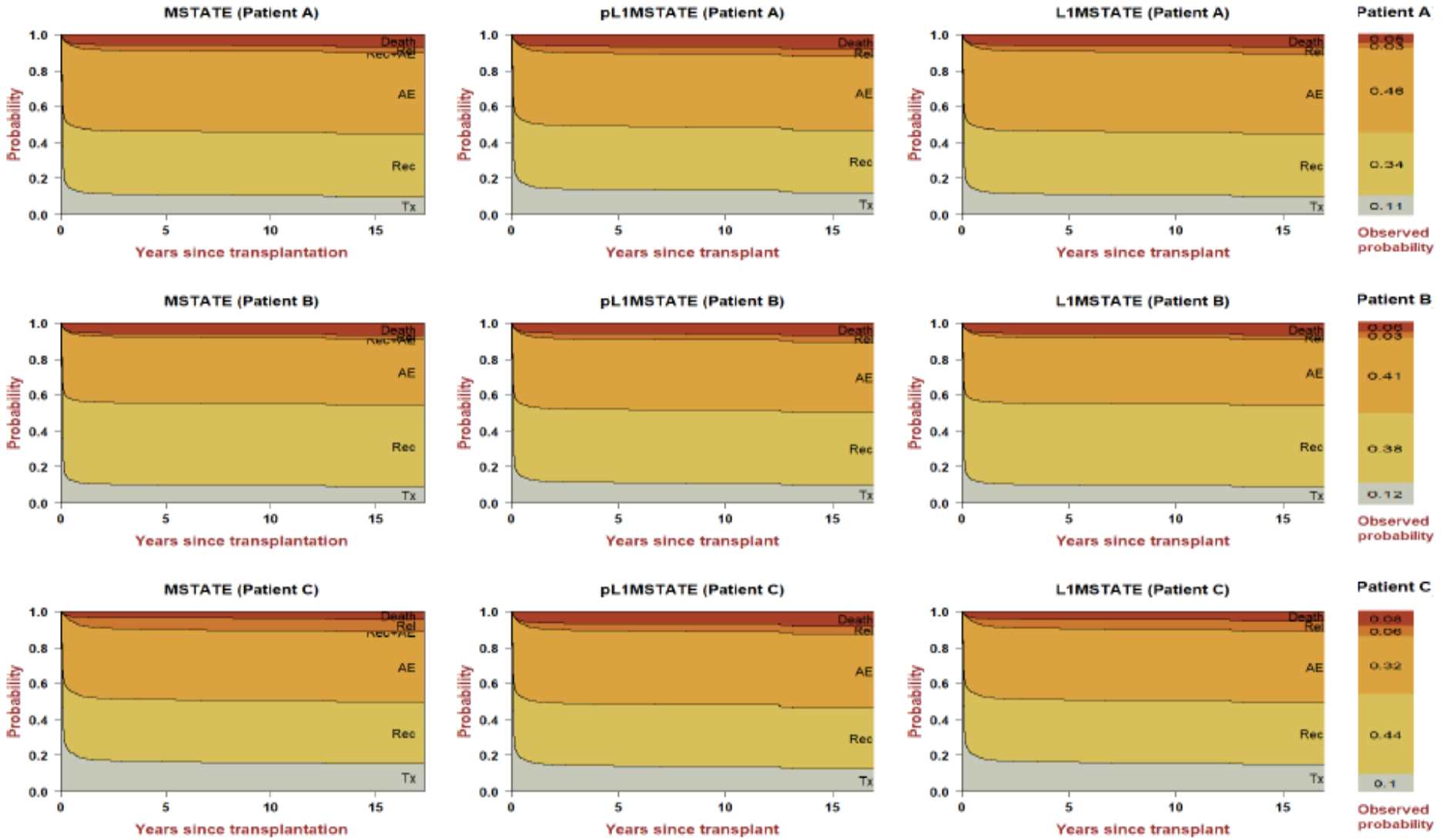


Fig. 6: Estimates of stacked prediction transition probabilities from $t = 0$ for patients A, B, and C using two models. L1MSTATE, L1 regularized multi-state model using the first cross-validation method; pL1MSTATE, L1-regularized multi-state model using the penalized cross-validation method; MSTATE, multi-state model.

Figure 7 shows the results of the Nelson-Aalen estimates of the four transitions starting from the transplant state for two patients A and D. There is a significant difference of the cumulative hazard rates of the first transition (from transplant state to recovery state) between the two patients. In other words, the year of transplant has significant effect upon the cumulative hazard function of the first transition: if patient did the transplant in 1995-1998, their cumulative hazard rate to recovery state is higher if they did in 1990-1994. The results of the predicted transition probabilities starting from the transplant state at starting computation time 0 of two patients in Figure 8 also show the strong effects of the year of transplant on the first transition probability. Note that it also shows the ability of risk factors (year of transplant) in discriminating patients who will have higher transferring risk (higher cumulative hazard and transition probability) by certain time (starting study time) from certain state (transplant).

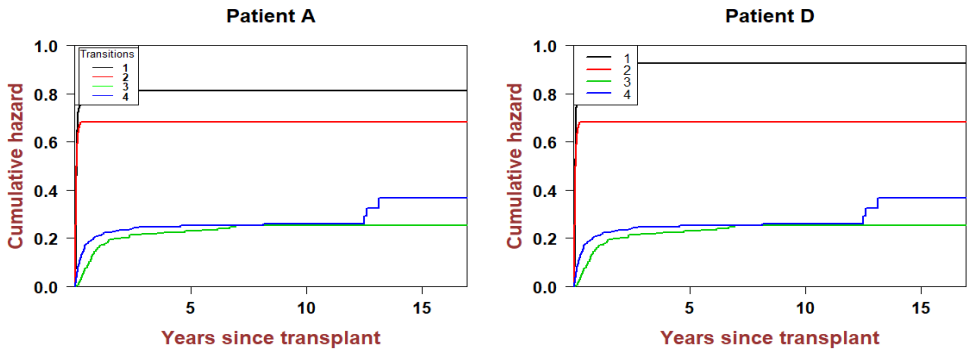


Fig. 7: Estimated cumulative hazard rates for patient A and patient D.

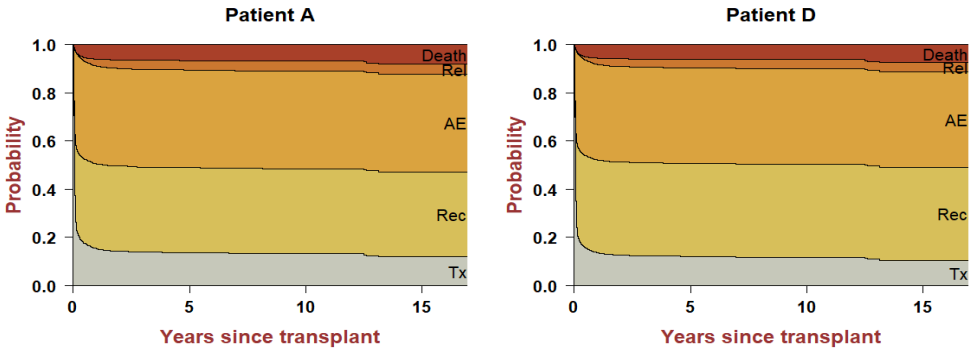


Fig. 8: Estimates of stacked prediction transition probabilities from $t = 0$ for patients A and D.

Table 14: The frequencies and proportions of the number of observed transitions from the transplant state of patient D after 0 and 100 days. The numbers in parentheses are proportions.

Days since transplant		Tx	Rec	AE	Rec+AE	Rel	Death	No event	Total
$t = 0$	Tx	0 (0)	94 (0.40)	85 (0.36)	0 (0)	4 (0.02)	15 (0.06)	5 (0.15)	233
$t = 100$	Tx	0 (0)	2 (0.04)	0 (0)	0 (0)	3 (0.07)	5 (0.11)	35 (0.78)	45

Although all the transition probabilities presented above are predicted at starting times 0, our '**L1mstate**' package also allows to compute the predicted transition probabilities at different starting computation times. For example, we can choose the starting computation times are 100 days since transplant to compute the predicted transition probabilities of patient D.

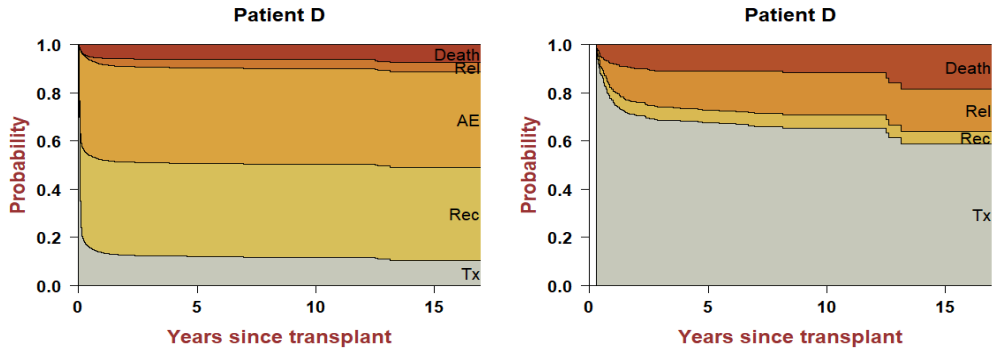


Fig. 9: Estimates of stacked prediction transition probabilities of patient D from $t = 0$ and $t = 100$ days since transplant.

Results in Figure 9 show the considerable changes of the distributions of the state probabilities: the probability of the transplant state increases substantially, and the probabilities of the relapse and death states also increase. In other words, if patient D can survive through the transplant state during the first 100 days, the chance that they may stay at the current state increases. Since the risk factors are assumed time-constant, this phenomenon may imply the effects of the risk factors upon the transition probabilities change over time or the sojourn time that patient D spent in the transplant state also affects upon the predicted transition probabilities.

6 Discussion

It is worthy of mentioning that we tried to apply the fused-lasso multi-state models method [38] using their R package '**penMSM**', but we could not obtain results due to the huge computation cost. For example, on the one hand, running one case (medium effects setting, $\rho = 0$, sample size of 350, and 20 λ values) with 20 replications took 24 hours to get the AUC values. It was run on a Dell Inspiron 15 computer (Intel Core i5-5200U 2.2GHz, 8GB RAM). On the other hand, our simulation studies included in total 228 cases with 100 replications for each case to get the final AUC values. Hence, the required computation time is way too much. However, we managed to obtain the results of TPRs, FPRs and AUC values without replication. We implemented the '**penMSM**' package by setting $\lambda_2 = 0$, and using 20 values of λ_1. To compute the TPR and FPR values, we used the Akaike Information Criterion (AIC) to select the optimal penalty parameters. To compute the AUC values, we calculated 20 pairs of TPRs and FPRs to construct ROC curve, then compute area under a ROC curve. The method was not able to perform automated variable selection since the results did not include zeros, so we rounded up the results to the 2^{nd} digit.

The results in Table 15 show that when $N = 250$ and $N = 450$, the fused-lasso multi-state model produces worse results with lower TPRs and higher FPRs comparing with L1MSTATE model and in many cases, its FPRs are higher than TPRs.

Table 15: Model selection results of Example I using the fused-lasso multi-state model approach.

N	ρ	Small effects		Medium effects		Large effects		Mixed effects	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
50	0	0.38	0.29	0.77	0.43	0.69	0.71	0.54	0.57
	0.2	0.69	0.5	0.69	0.64	0.69	0.79	0.54	0.64
	0.5	0.46	0.50	0.69	0.50	0.85	0.57	0.46	0.64
250	0	0.69	0.71	0.77	0.64	0.85	0.71	0.69	0.57
	0.2	0.62	0.71	0.62	0.79	0.92	0.64	0.62	0.64
	0.5	0.62	0.71	0.92	0.86	0.92	0.64	0.85	0.64
450	0	0.92	0.50	0.92	0.93	0.85	0.79	0.85	0.64
	0.2	0.77	0.86	1.00	0.93	0.85	0.86	0.62	0.79
	0.5	0.62	0.42	0.77	0.86	1.00	0.86	0.92	0.86

TPR: true positive rate; FPR: false positive rate.

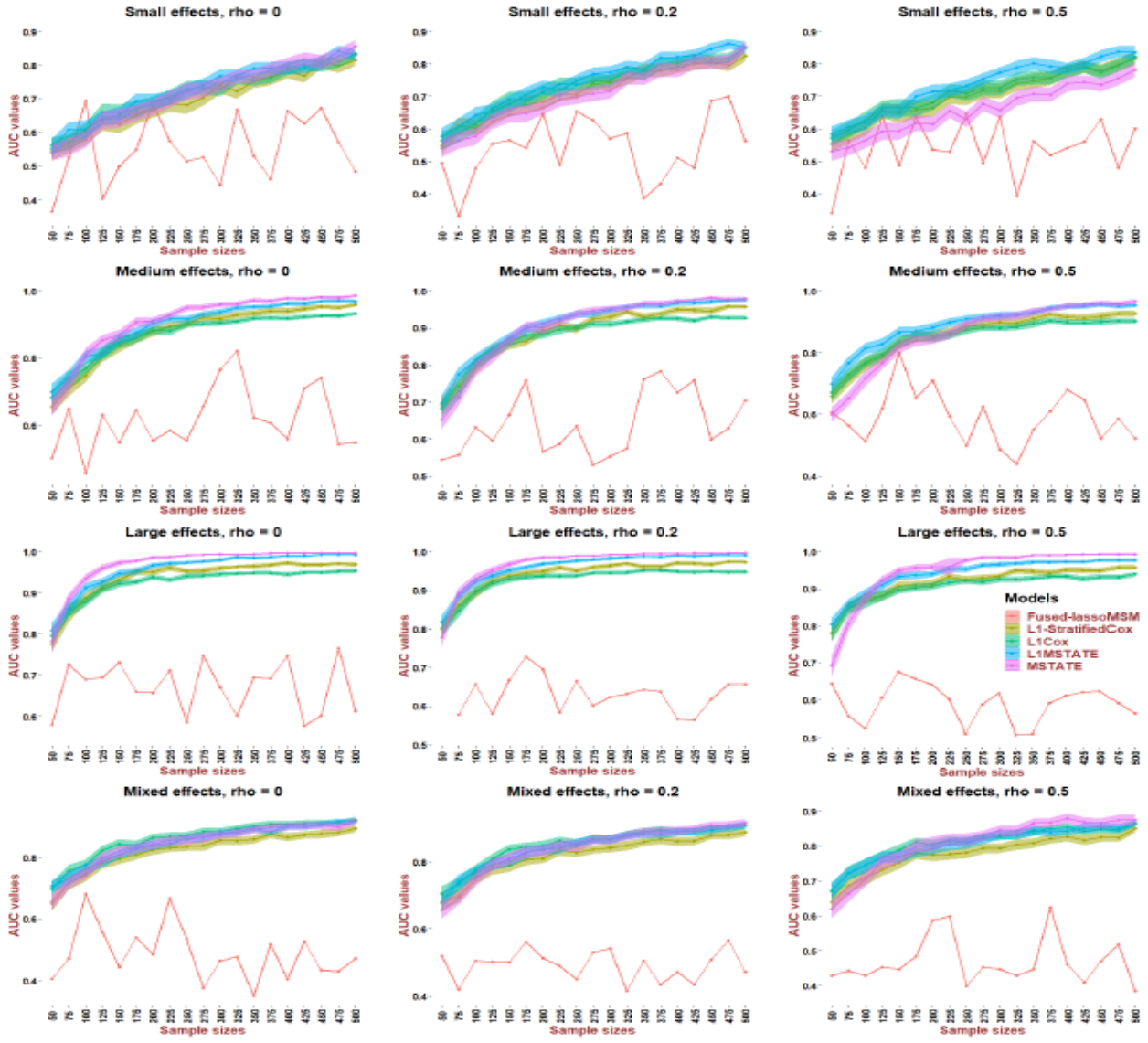


Fig. 10: AUC values of Example I for different sample sizes in different settings: 100 replications for each of first four models, no replication for the fused-lasso multi-state model (penMSM).

Figure 10 shows that the fused-lasso multi-state model produces the worst performance in term of the AUC values. In other words, the fused-lasso multi-state model method does not do variable selection while our L1MSTATE method is developed for this purpose. Our simulation studies showed that using the penalized cross-validation method to select the optimal tuning parameter produced more sparse models than using the first cross-validation method in our L1-regularized multi-state models. In some cases, however, it tends to suppress too much. For example, in EBMT study, the results from Table 11 show that using the penalized cross-validation method suppresses almost all of the risk factors in some *rare* transitions such as transitions 3, 4 and 9. Hence, the penalized cross-validation method should be used with caution in real world applications.

7 Conclusions and future works

We propose the L1-regularized multi-state model framework for simultaneous parameter estimation and variable selection using the L1-regularized partial likelihood approach. We devise the one-step coordinate descent algorithm and use a local quadratic approximation of the log-partial likelihood to solve the corresponding optimization problem, which can offer significant improvement on the computational efficiency. Our proposed method demonstrates the state-of-the-art performance in terms of identifying the significant risk factors comparing with the existing regularized multi-state models in simulation studies. It also performs better at doing variable selection and predicting the transition probabilities in cases with small sample sizes comparing with the un-regularized approach in simulation and real-world cases.

Here we focus on specifying a Cox model for each transition, but our approach can be easily extended to other types of models for each transition. Our L1-regularized multi-state models can be applied to competing risks data in the cause-specific hazards models setting. If other models such as the subdistribution hazard model or competing risks quantile regression model are used to analyze competing risks data, the literature on variable selection using regularized approaches can be found in several papers [20], [4], [18], [39], [3], [28], [17].

In this paper, we also assumed that the coefficients are constant over time, but it is common in longitudinal studies to collect information of the same risk factors at follow-up visits. Therefore, it may be beneficial to utilize the time-dependent risk factors. One approach is to use piecewise constant coefficients that allows time-varying risk factors, but it may require totally different techniques for model inference. Furthermore, its model may be complicated and hard to interpret. Another approach is to use joint models of time-to-event and longitudinal data which will be future research topics.

In applying our L1-regularized multi-state models to EBMT dataset, we used dummy variables for two categorical risk factors with three levels, but our current approach cannot guarantee these dummy variables enter or leave the model together. A more appropriate way to handle categorical risk factors is through group lasso penalty [45]. In addition, the analyses of the risk factor effects upon the cumulative hazard functions and the transition probabilities suggest two extensions should be considered: one is about the time-dependent effects of risk factors in multi-state models and the other is the semi-Markov multi-state models. These would be our future directions to extend our method.

Acknowledgments

This work was partially supported by the National Science Foundation (NSF)– Division of Communication and Computing Foundations (CCF) awards #1718513, #1715027, #1714136 and the JDRF award #2-SRA-2018-513-S-B.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Aalen, O.O., Borgan, O., Gjessing, H.K.: Survival and event history analysis. A process point of view. Springer, New York, NY (2008). ISBN 978-0-387-20287-7
2. Aalen, O.O., Johansen, S.: Empirical transition matrix for nonhomogeneous Markov-chains based on censored observations. Scandinavian Journal of Statistics **5**, 141–150 (1978)
3. Ahn, K., Banerijee, A., Sahr, N., Kim, S.: Group and within-group variable selection for competing risks data. Lifetime Data Analysis **24**(3), 407–424 (2018)
4. Ambrogi, F., Scheike, T.: Penalized estimation for competing risks regression with applications to high-dimensional covariates. Biostatistics **17**(4), 708—-721 (2016)
5. Andersen, P.K.: Multistate models in survival analysis: a study of nephropathy and mortality in diabetes. Statistics in Medicine **7**(6), 661–670 (1988)
6. Andersen, P.K., Borgan, O., Gill, R.D., Keiding, N.: Statistical models based on counting processes. Springer, New York, NY (1993). ISBN 978-1-4612-4348-9
7. Andersen, P.K., Hansen, L.S., Keiding, N.: Assessing the influence of reversible disease indicators on survival. Statistics in Medicine **10**, 1061–1067 (1991)
8. Andersen, P.K., Keiding, N.: Multi-state models for event history analysis. Statistical Methods Medical Research **11**(2), 91–115 (2002)
9. Breslow, N.E.: Discussion of the paper by D.R.Cox. Journal of the Royal Statistical Society Series B **34**, 216–217 (1972)
10. Chen, H.H., Duffy, S.W., Tabar, L.: An arbitrary Lagrangian-Eulerian computing method for all flow speeds. J Comput Phys **14**(3), 227–253 (1974)
11. Commenges, D., Joly, P., Letenneur, L., Dartigues, J.F.: Incidence and mortality of Alzheimer's disease or dementia using an illness-death model. Statistics in Medicine **23**, 199–210 (2004)
12. Cox, D.R.: Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological) **34**(1), 187–220 (1972)
13. deWreede, L.C.: mstate: An r package for the analysis of competing risks and multi-state models. Journal of Statistical Software **38**(7), 53–66 (2011)
14. deWreede, L.C., Fiocco, M., Putter, H.: The mstate Package for Estimation and Prediction in Non- and Semi-Parametric Multi-State and Competing Risks Models. Computer Methods and Programs in Biomedicine **99**(3), 261–74 (2010)

15. Duffy, S.W., Chen, H.H.: Estimation of mean sojourn time in breast cancer screening using a Markov chain model of entry to and exit from preclinical detectable phase. Statistics in Medicine **14**, 1531–1543 (1995)
16. Fawcett, T.: An introduction to roc analysis. Pattern Recognition Letters **27**, 861–874 (2006)
17. Fu, Z., Ma, S., Lin, H., Parikh, C., Zhou, B.: Penalized variable selection for multi-center competing risks data. Statistics in Biosciences **9**, 379–405 (2017)
18. Fu, Z., Parikh, C., Zhou, B.: Penalized variable selection in competing risks regression. Lifetime Data Analysis **23**, 353–376 (2017)
19. Gentleman, R.C., Lawless, J.F., Lindsey, J.C., Yan, P.: Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. Statistics in Medicine **13**(3), 805–821 (1994)
20. Ha, I., Lee, M., Oh, S., Jeong, J., Sylvester, R., Lee, Y.: Variable selection in subdistribution hazard frailty models with competing risks data. Statistics in Medicine **30**(26), 4590—-4604 (2014)
21. Hastie, T., Tibshirani, R.: Generalized Additive Models. Chapman and Hall, London (1990). ISBN 9780412343902
22. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Prediction, Inference and Data Mining. Springer, New York, NY (2009). ISBN 978-0-387-84858-7
23. Huang, S., Hu, C., Bell, M., Billheimer, D., Guerra, S., Roe, D., Vasquez, M., Bedrick, E.: Regularized continuous-time markov model via elastic net. Biometrics **74**(3), 1045– 1054 (2018)
24. Jackson, C.H., Sharples, L.D., Thompson, S.G., Duffy, S.W., Couto, E.: Multistate Markov models for disease progression with classification error. Journal of the Royal Statistical Society Series D - The Statistician **52**(2), 193–209 (2003)
25. Johansen, S.: An Extension of Cox's Regression Model. International Statistical Review **51**(2), 165–174 (1983)
26. Kalbfleisch, J., Lawless, J.F.: The analysis of panel data under a Markov assumption. Journal of American Statistical Association **80**(392), 863—871 (1985)
27. Kay, R.: A Markov model for analyzing cancer markers and disease states in survival studies. Biometrics **42**, 855–865 (1986)
28. Kim, S., Ahn, K.: Bi-level variable selection for case-cohort studies with group variables. Statistical Methods in Medical Research **28**(10-11), 3404–3414 (2019)
29. Kirby, A.J.: Statistical modelling for the precursors of cervical cancer. Tech. Rep. Thesis (Ph.D.), University of Cambridge, Cambridge, England, United Kingdom (1991)
30. Klotz, J.H., Sharples, L.D.: Estimation for a Markov heart transplant model. The Statistician **43**(3), 431–436 (1994)
31. Longini, I.M., Clark, W.S., Byers RA HA G.F., Hethcote, H.W.: Statistical analysis of the stages of HIV infection using a Markov model. Statistics in Medicine **8**, 851–843 (1989)
32. Mairal, J., Yu, B.: Complexity analysis of the lasso regularization path. Proceedings of the 29th. International Conference on Machine Learning, Edinburgh, Scotland, UK (2012)
33. Marshall, G., Jones, R.H.: Multi-state Markov models and diabetic retinopathy. Statistics in Medicine **14**(18), 1975–83 (1995)
34. Meier, L., vandeGeer, S., Buhlmann, P.: The group lasso for logistic regression. Journal of the Royal Statistical Society Series B **70**(1), 53–71 (2007)
35. Oelker, M., Tutz, G.: A uniform framework for the combination of penalties in generalized structured models. Advances in Data Analysis and Classification **11**(1), 97–120 (2017)
36. Perez-Ocon, R., Ruiz-Castro, J., Gamiz-Perez, M.: Non-homogeneous Markov models in the analysis of survival after breast cancer. Journal of the Royal Statistical Society Series C-Applied Statistics **50**, 111–124 (2001)
37. Putter, H., Fiocco, M., Geskus, R.B.: Tutorial in biostatistics: Competing risks and multistate models. Statistics in Medicine **26**, 2389–2430 (2007)
38. Reulen, H., Kneib, T.: Structured fusion lasso penalized multi-state models. Statistics in Medicine **35**(25), 4637—4659 (2016)
39. Saadati, M., Beyersmann, J., Kopp-Schneider, A., Benner, A.: Prediction accuracy and variable selection for penalized cause-specific hazards models. Biometrical Journal **60**(2), 288–306 (2018)
40. Sharples, L.D.: Use of the Gibbs sampler to estimate transition rates between grades of coronary disease following cardiac transplantation. Statistics in Medicine **12**, 1155–1169 (1993)
41. Simon, N.: Regularization paths for cox's proportional hazards model via coordinate descent. Journal of Statistical Software **39**(5), 53–66 (2012)

42. Ternes, N., Rotolo, F., Michiels, S.: Empirical extensions of the lasso penalty to reduce the false discovery rate in high-dimensional Cox regression models. Statistics in Medicine **35**(15), 2561–73 (2016)
43. Tibshirani, R.: The lasso method for variable selection in the cox model. Statistics in Medicine **16**(4), 385–395 (1996)
44. Verweij, P.J., Houwelingen, H.C.: Cross-validation in survival analysis. Statistics in Medicine **12**(24), 385–395 (1993)
45. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society Series B (Methodological) **68**(1), 49–67 (2006)