RESEARCH ARTICLE | JUNE 07 2024

Estimation of the transition probabilities conditional on covariates with repeated measures: A joint modeling approach

Gustavo Soutinho; Luís Meira-Machado

AIP Conf. Proc. 3094, 450004 (2024)
https://doi.org/10.1063/5.0211213

Estimation of the Transition Probabilities Conditional on Covariates with Repeated Measures: a Joint Modeling Approach

Gustavo Soutinho$^{1,a)}$ and Luís Meira-Machado$^{2,b)}$

^{1}*EPIUnit - University of Porto, Rua das Taipas 135, 4050-600 Porto, Portugal*
^{2}*Centre of Mathematics, University of Minho, Campus de Azurém, 4800 - 058 Guimarães, Portugal*

$^{a)}$Corresponding author: gdsoutinho@gmail.com
$^{b)}$lmachado@math.uminho.pt

Abstract. In recent years, there has been a significant urge of interest in longitudinal and survival data modeling. This approach holds particular significance in cancer research, where it enables the evaluation of how longitudinal markers influence the event of interest. This paper aims to introduce practical estimation techniques for transition probabilities, conditional on observed covariates with repeated measurements. This innovation allows us to incorporate the trajectory of longitudinal outcomes into regression models by accommodating time-varying covariates for each individual. The results presented in this study confirm the superior efficiency of the proposed methods, which merge existing approaches for joint modeling of longitudinal and survival data with the landmark approach for estimating transition probabilities. These methods outperform approaches that do not fully account the information provided by longitudinal covariate measurements.

INTRODUCTION

A multi-state model is a mathematical framework for characterizing continuous-time stochastic processes, particularly useful for analyzing complex event histories involving multiple events. In biomedical research, these models often used to monitor and analyze the progression of patients through various disease stages, such as those encountered in cancer disease or HIV infection. The states can be defined based on clinical manifestations, biological indicators, or non-fatal complications arising during the course of the illness (Andersen and Keiding, 2002; Rizopoulos, 2012; Proust-Lima and Taylor, 2009).

In the context of multi-state modeling, one significant objective, yet underexplored, is the estimation of transition probabilities conditional on a continuous covariate repeatedly observed over time. This longitudinal data collection involves gathering data at multiple time points for each study participant and is characterized by the interdependence of repeated observations over time within each participant. A common and simplistic approach is to model these processes separately, utilizing linear mixed effects models for the longitudinal model and Cox regression models for time-to-event analysis (Cox, 1972). However, this practice is generally discouraged in practical applications due to its tendency to yield biased effect size estimates, primarily attributed to the correlation between the two outcomes (Ibrahim, Chu and Chen, 2010; Hickey *et al.*, 2018).

This paper revisits the challenge of estimating transition probabilities in (potentially) non-Markov models. However, unlike prior contributions (Meira-Machado *et al.*, 2006; Meira-Machado, de Uña-Álvarez and Datta, 2015), our focus lies in estimating these probabilities while considering longitudinal biomarkers. To achieve this, we employ the landmark approach, as introduced by de Uña-Álvarez and Meira-Machado (2015), in combination with the joint modeling technique (Rizopoulos, 2012). The resulting model is constructed by integrating two sub-models: one for the longitudinal aspect (e.g., a linear mixed effects model) and another for time-to-event analysis (e.g., a proportional hazards model) for each transition intensity. Comprehensive discussions regarding the foundational concepts extending joint modeling to multi-state models can be found in Ferrer *et al.* (2016).

International Conference of Numerical Analysis and Applied Mathematics
AIP Conf. Proc. 3094, 450004-1–450004-4; https://doi.org/10.1063/5.0211213

METHODS

Notation and preliminaries

The illness-death model stands out as one of the most prevalent multi-state models in the literature. In its irreversible form, individuals initiate in an initial state (referred to as State 0 in this paper). From there, they progress to either an intermediate transient state (State 1) or transition to an absorving 'dead' state (State 2). Once individuals reach the 'diseased' state, there is no possibility of recovery, and they ultimately transition to the 'dead' state. This process is defined by the joint distribution of $(Z,\ T)$, where Z represents the duration spent in State 0, and T denotes the total survival time of the process. In practice, the observation of these two variables is influenced by right censoring issues. This censoring can arise due to various factors such as time limitations in follow-up, cases lost to follow-up, and more. To account for this, we introduce a censoring variable, denoted as C, which takes on binary values to indicate the presence or absence of the event. Importantly, we assume that this censoring variable is independent of the underlying process.

Joint modeling specification

The joint modeling approach for multi-state models can be described through a linear mixed-effects model and a survival sub-model for each transition. The longitudinal sub-model adheres to Gaussian assumptions, where the observed measurement Y_{ij} at time t_{ij} is expressed as:

$$Y_{ij} = X_i\left(t_{ij}\right)^T \beta + Z_i\left(t_{ij}\right)^T b_i + \varepsilon_{ij}.$$

Here $X_i\left(t_{ij}\right)$ and $Z_i\left(t_{ij}\right)$ denote the vectors of time-dependent covariates for the individual, and b_i is the vector of random effects, with b_i following a normal distribution $N(0, \Sigma)$. The parameter β represents a fixed vector, and ε_i follows a normal distribution $N\left(0, \sigma^2 I_{n_i}\right)$, where n_i represents the number of longitudinal measurements for each individual (Ferrer *et al.*, 2016).

The survival outcome, occurring at time t and transitioning from state h to state k (where h and k belong to the finite state space S), is represented by a Cox sub-model, which is formulated as follows:

$$\lambda^i_{hk}\left(t|b_i\right) = \lambda_{hk,0}\left(t\right) \exp\left\{X^{ST}_{hk,i}\gamma_{hk} + W_{hk,i}\left(b_i, t\right)\eta_{hk}\right\}$$

Here, $\lambda_{hk,0}\left(.\right)$ denotes a parametric baseline intensity, which can follow distributions such as Weibull, exponential, or piecewise constant. γ_{hk} represents a vector of regression coefficients for the covariates $X^S_{hk,i}$. The parameter η_{hk} quantifies the influence of the longitudinal outcome on the time-to-event for the transition from state h to state k. The multivariate function $W_{hk,i}\left(b_i, t\right)$ characterizes the dependency structure between the longitudinal and multi-state processes and represents the unobserved true value of the longitudinal outcome for patient i at time t.

Estimation

The parameters for the joint models are estimated using Maximum Likelihood Estimation (MLE), where the log-likelihood function is maximized. In cases of slow convergence, this maximization is achieved through an EM algorithm, often coupled with a quasi-Newton algorithm. As previously mentioned, the primary objective of this paper is to estimate conditional transition probabilities denoted as $p_{hj}(s, t \mid Y)$ (Meira-Machado *et al.*, 2015), where Y represents a covariate with longitudinal measurements, such as tumor markers measured at different time points. Employing the landmark approach, our interest lies in predicting the conditional transition probabilities for a new subject, denoted as i, possessing a set of longitudinal measurements $\widetilde{y}_i(v) = \{y_i(u), 0 \leq u \leq v\}$. One of these quantities is expressed by the conditional probability $p^i_{00}(s, t \mid Y_i) = P(Z > t|Z > s, y_i(s), S_0)$, where $t > s$, and $S_0 = \{i : \widetilde{Z}_i > s\}$ represents the landmark sample used to fit the joint model. Another intriguing quantity is the conditional probability $p^i_{11}(s, t \mid Y_i) = P(T > t|Z \leq s, T > s, y_i(s), S_1)$, with $S_1 = \{i : \widetilde{Z}_i \leq s, \widetilde{T}_i > s\}$. A Monte Carlo estimation of these probabilities can be obtained through a Bayesian formulation of the problem. For more detailed information, please refer to Rizopoulos (2012).

SIMULATION STUDY

To generate the joint model data, we used 1000 replicas, each one with sample sizes of 400 observations, through the following two sub-models, $Y_{ij} = \beta_0 + \beta_1 \times t_{ij} + b_{i0} + b_{i1} \times t_{ij} + \varepsilon_{ij}$ and $\lambda^i_{hk}\left(t|b_i\right) = \lambda_{hk,0}\left(t\right) \exp\{\gamma_{hk} + W_{hk,i}\left(b_i, t\right)\eta_{hk}\}$. Here

$h \in \{0,1\}$ and $k \in \{0,1,2\}$ correspond to the three states $\{0,1,2\}$, from which we denote the transition 1 (between State 0 and State 1), transition 2 (between State 0 and State 2) and transition 3 (between State 1 and State 2). For each of the observations were simulated repeated measures using the same times $t_{ij} = 0.33, 0.66, \cdots, 16.50$ and longitudinal sub-models parameters generated from normal distributions, $b_i \sim N\left((0,0)^T, \begin{bmatrix} 20 & 0.2 \\ 0.2 & 0.02 \end{bmatrix}\right)$ and $\varepsilon_i \sim N(0, 18)$. Finally, the survival sub-models were simulated considering that $\lambda_{hk,0}(.) \sim \exp(\Lambda)$, with parameters $\Lambda \in \{3, 1.7, 0.5\}$, and $\eta_{hk} = \{-0.7, -0.7, -0.6\}$, respectively, for transitions 1, 2 and 3. For all three Cox sub-models, $\gamma_{hk} = \{2\}$. For chosing the previous values for parameters was taken into consideration percentage of events and censoring. The vector of true transition times, $\left(T^*_{i,01}, T^*_{i,02}, T^*_{i,12}\right)$ was generated following the procedures outlined in Beyersmann et al. (2011). The available times $T_{i,hj}$ are limited to their censored versions, and the longitudinal measurements, generated from the linear mixed sub-model, were truncated at $T_{i,01}$, which corresponds to the first observed time of the multi-state process.

Results

The transition probabilities for the three different methods: the Landmark approach (`LM`), Breslow's method (`BRES`), and the Joint Modeling-Landmark estimator (`JMLM`), were computed using Monte Carlo simulations involving 1000 replications with 400 individuals. In each replication, a subset of eight individuals was retained to assess the impact of longitudinal marker trends (including decreasing, constant, increasing, and random patterns) on the estimation of transition probabilities. The results consistently demonstrate that the `JMLM` estimator outperforms the other two methods for all probabilities $p_{00}(8, t \mid Y)$, regardless of the trend in the longitudinal marker. The boxplots for the `BRES` and `LM` estimators exhibit a systematic bias, suggesting that these methods may not be suitable for accurately capturing the evolution of individuals' repeated markers (see Figure 1).

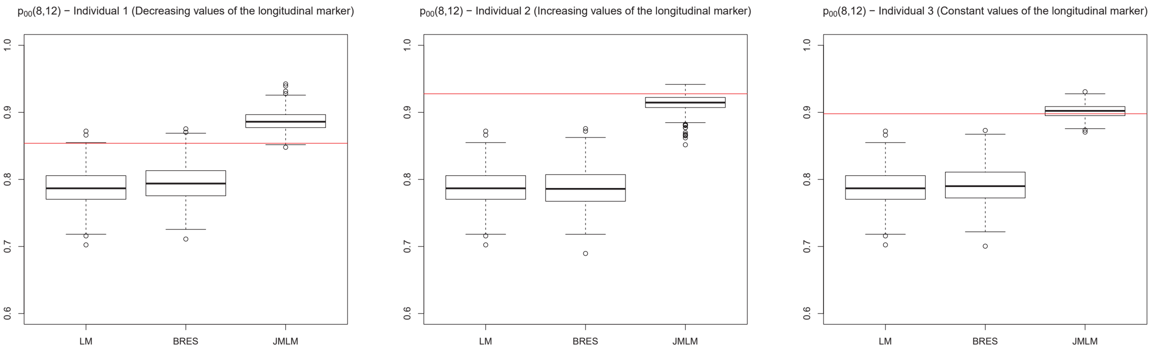

FIGURE 1. Boxplots of the M = 1000 estimates of the transition probabilities.

Upon comparing the estimated values of $p_{00}(8, 12 \mid Y)$ and $p_{00}(8, 18 \mid Y)$ between individual 1 and individual 2, we can infer that an increasing trend in the longitudinal marker corresponds to a higher true value. These results also highlight the `JMLM` estimator's capacity to accurately represent the longitudinal marker's evolution. For example, when considering $\hat{p}_{00}(8, 12)$ for individual 2, who exhibits an increasing trend in the longitudinal marker, the Breslow estimator accounts for the higher value, resulting in a decrease in transition probabilities compared to the `LM` estimator. However, the influence of past repeated measurements leads to an increase in the `JMLM` estimation, aligning it more closely with the true values.

APPLICATION TO A DATA SET

In this section, we assess the performance of the proposed `JMLM` estimators using two datasets containing longitudinal and survival data, which were previously analyzed by Ferrer *et al.* (2016) (refer to their Section 4 for detailed information). Specifically, our goal is to compare the `JMLM` estimator with estimators based on proportional hazards model, fitted marginally to the corresponding transition, with the baseline hazard function estimated using Breslow's method (referred to as `BRES`). For comprehensive analysis, we also include the original landmark estimator, denoted as `LM`, which does not account for the covariate's effect. To assess the behavior of these three estimators in this section, we evaluate the conditional transition probabilities at a fixed value of $s = 8$. This value is used to construct two landmark datasets based on individuals in State 0 ($\mathcal{S}_0 = \{i : \widetilde{Z}_i > 8\}$) and State 1 ($\mathcal{S}_1 = \{i : \widetilde{Z}_i \leq 8, \widetilde{T}_i > 8\}$). The first subsample of data is employed to estimate the conditional transition probabilities $p^i_{00}(8, t|Y_i(t))$, while the second subsample is utilized to estimate $p^i_{11}(8, t|Y_i(t))$. Figure 2 illustrates the utility of the proposed methods, highlighting the greater variability of the `BRES` estimator, particularly for longer lag times ($t - s$). Moreover, the proposed `JMLM` estimator demonstrates its ability to capture the impact of the longitudinal marker's progression on the transition probability estimates.

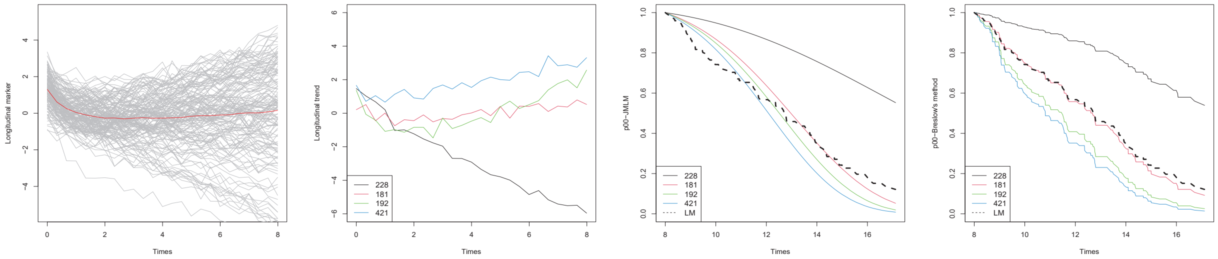

FIGURE 2. Longitudinal values for the subsets S_0 for all and four chosen subjects (First two figures to the left, respectively). Last two figures to the right show the estimated transition probabilities $\hat{p}^i_{00}(8, t \mid Y_i)$ using the JMLM (third) estimator and the BRES estimator (fourth) for the selected individuals that belong to the subset S_0.

DISCUSSION

The results from the simulation studies consistently validated the strong performance of the proposed estimator, which yielded precise estimates of transition probabilities. Furthermore, the proposed method exhibited a higher sensitivity in capturing the longitudinal measures' evolution when contrasted with the Breslow-based method, which relies solely on a single covariate value.

ACKNOWLEDGMENTS

This work received financial support from the Portuguese Foundation for Science and Technology, references UIDB/00013/2020, UIDP/00013/2020, EXPL/MAT-STA/0956/2021 and PD/BD/142887/2018.

REFERENCES

[1] P. K. Andersen and N. Keiding, Statistical Methods in Medical Research **11**, 91–115 (2002).
[2] D. Rizopoulos, *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R* (Chapman and Hall/CRC, New York, 2012).
[3] C. Proust-Lima and J. Taylor, Biostatistics **10**, 535–549 (2009).
[4] J. Ibrahim, H. Chu, and L. Chen, Journal of Clinical Oncology **28**, 2796–2801 (2010).
[5] G. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona, The International Journal of Biostatistics **14**, 1–19 (2009).
[6] L. Meira-Machado, J. de Uña-Álvarez, C. Cadarso-Suárez, and P. K. Andersen, Statistical Methods in Medical Research **18**, 195–222 (2009).
[7] J. d. U.-l. L. Meira-Machado and S. Datta, Computational Statistics **30**, 377–397 (2015).
[8] L. Ferrer, V. Rondeau, J. Dignam, T. Pickles, H. Jacqmin-Gadda, and C. Proust-Lima, Statistics in Medicine **35**, 3933–3948 (2016).
[9] J. de Uña-Álvarez and L. Meira-Machado, Biometrics **71**, 364–375 (2015).