Direct modeling of regression effects for transition probabilities in the progressive illness-death model

Leyla Azarang, Thomas Scheike, Jacobo De uña-Álvarez

Statistics
in Medicine

Research Article

Received 26 November 2015, Accepted 18 January 2017 Published online 26 February 2017 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7245

Direct modeling of regression effects for transition probabilities in the progressive illness–death model

Leyla Azarang,[a*†] Thomas Scheike[b] and Jacobo de Uña-Álvarez[c,d]

In this work, we present direct regression analysis for the transition probabilities in the possibly non-Markov progressive illness–death model. The method is based on binomial regression, where the response is the indicator of the occupancy for the given state along time. Randomly weighted score equations that are able to remove the bias due to censoring are introduced. By solving these equations, one can estimate the possibly time-varying regression coefficients, which have an immediate interpretation as covariate effects on the transition probabilities. The performance of the proposed estimator is investigated through simulations. We apply the method to data from the Registry of Systematic Lupus Erythematosus RELESSER, a multicenter registry created by the Spanish Society of Rheumatology. Specifically, we investigate the effect of age at Lupus diagnosis, sex, and ethnicity on the probability of damage and death along time. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: binomial regression; censored data; illness–death model; transition probabilities

1. Introduction

The three states progressive illness–death model (also known as disability model) is useful for investigating the occurrence of an intermediate state in chronic diseases as well as death and is relevant for irreversible diseases where recovery is impossible. The model involves three states: 'Healthy' (state 1), 'Diseased' (state 2), and 'Dead' (state 3), and three possible transitions among them: 1 → 2, 2→3, and 1 → 3 (Figure 1). In this model, states 1 and 2 are transient, while state 3 is absorbing. The model may be used, for example, to describe the disease process in cancer studies. Also, in epidemiology, it is applied to investigate both incidence of a disease and death. Another example is the study of systemic lupus erythematosus (SLE) disease in Section 4, where 'No damage', 'Damage', and 'Dead' are identified as relevant states (Figure 4).

Often, it is of interest to estimate the transition probabilities for the illness–death model, because they allow for long-time predictions [1]. For the lupus data in Section 4, these transition probabilities serve to evaluate the damage-free and total survival probabilities, among other curves of interest. The standard nonparametric approach to estimate the transition probabilities under the Markov assumption is the time honored Aalen–Johansen estimator [2]. Because the Markov assumption ignores disease history, it might be inappropriate in many settings. Meira-Machado *et al.* [3] introduced for the first time non-Markov nonparametric estimators for the transition probabilities for the progressive illness–death model. Similar approaches were developed by Allignol *et al.* [4], Titman [5], and de Uña-Álvarez and Meira-Machado [6]. In particular, the latter paper introduces a simple estimator for the transition probability matrix, which depends on the Kaplan–Meier estimators computed for different event times and specific subsamples. In

[a]*Aix-Marseille Univ, INSERM, IRD, SESSTIM, Sciences Economiques et Sociales de la Santé et Traitement de l'Information Médicale, Marseille, France*
[b]*Department of Public Health, University of Copenhagen, Copenhagen, Denmark*
[c]*Statistical Inference, Decision and Operations Research (SiDOR) Group, University of Vigo, Spain*
[d]*Department of Statistics and Operations Research Centre for Biomedical Research (CINBIO), University of Vigo, Spain*
[*]*Correspondence to: Leyla Azarang, Aix-Marseille Univ, INSERM, IRD, SESSTIM, Sciences Economiques et Sociales de la Santé et Traitement de l'Information Médicale, Marseille, France.*
[†]*E-mail: leyla.azarang@inserm.fr*

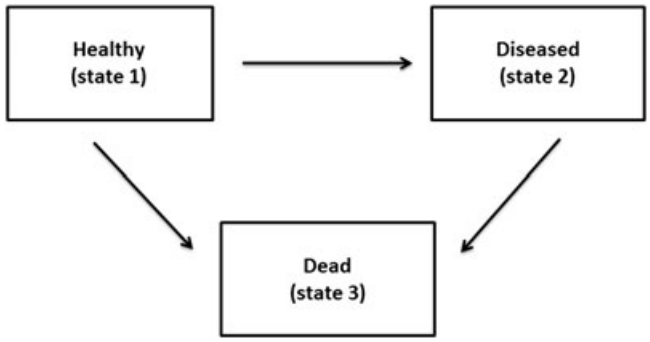

Figure 1. Progressive illness–death model.

a sense, the method we introduce in this paper to estimate regression effects for transition probabilities is a more general case of such simple approach.

We consider baseline covariates. In the presence of censoring, Aalen *et al.* [7] provided estimators for the transition rates based on an additive model and then combined the estimated rates conditionally on the covariates into appropriate conditional transition probabilities. This approach is similar to Aalen's additive model for the cause-specific hazard in competing risks model, for which the Markov condition is always satisfied, and it does not allow for a direct estimation of covariate effects on the transition probabilities. For competing risks, Scheike *et al.* [8] showed that translating the effects on cause-specific hazards into effects on the cumulative incidence functions (which are particular transition probabilities) is difficult in general.

The subdistribution approach by Fine and Gray [9] provides estimators for the covariate effects on the cumulative incidence functions by solving the inverse probability of censoring weighted (IPCW) version of Cox-type score equations. Other existing techniques to directly estimate effects on transition probabilities are the pseudo-value approach (cf. [10, 11]) and the binomial regression approach based on IPCW score equations by Scheike *et al.* [8]. Both the binomial regression approach and the pseudo-value approach allow for a variety of link functions. However, the choice of the link function is important for the interpretation of the regression parameters [12]. Meira-Machado *et al.* [13] considered the estimation of transition probabilities conditioning on continuous covariates. Their approach is based on kernel smoothing, which can be applied to multiple covariates but suffers from the curse of dimensionality even in low dimension. Besides, it does not allow to incorporate categorical covariates, which often appear in biomedical practice.

In this paper, we address the problem of estimating the transition probabilities in a possibly non-Markov progressive illness–death model in the presence of covariates, using a binomial approach analogous to that in [8] for competing risks. We account for possible violations of the Markov property using a subsample idea used by Allignol *et al.* [4] in the setting without covariates. For this purpose, we proceed by restricting the sample to two subsamples depending on whether the transition is made from the initial state or intermediate state. In the first case, the subsample is the set of individuals observed in the initial state by a given time s and, in the second case, those observed in the intermediate state by that time. It is assumed that all the individuals are in the initial state by time zero. The proposed method can be applied for both continuous and categorical covariates and, because of its semi-parametric structure, it allows for the construction of accurate estimators regardless of the dimension of the covariate vector. Furthermore, the given semi-parametric approach allows for the interpretation of the covariate effects on transition probabilities in a simple way.

The rest of the paper is organized as follows. In Section 2, we describe the model in detail, and we introduce the new method, which is investigated through simulations in Section 3. In Section 4, we apply the method to data from the Registry of Systematic Lupus Erythematosus Patients of the Spanish Society of Rheumatology (RELESSER), a multicenter registry created by the Spanish Society of Rheumatology. Specifically, we investigate the effect of age at Lupus diagnosis, sex, and ethnicity on the probability of damage and death along time. In Section 5, a final discussion is given.

2. Model and estimators

The progressive illness–death model (Figure 1) is a three-state model that allows transitions in only one direction and which consists of an initial state (state 1), an intermediate, transient state (state 2), and a final absorbing state (state 3). All the individuals are assumed to be in the initial state at time zero. The sojourn time in the initial state and the total survival time are denoted by Z and T, respectively. In the case of a direct transition from the initial state to the final state, we have $T = Z$. Also, a vector of time-independent covariates $\boldsymbol{X}$ is available. We consider possibly right-censored data. Then, instead of Z and T, we observe $\tilde{Z} = \min(Z, C)$ and $\tilde{T} = \min(T, C)$, where C is the potential censoring time, which we assume to be independent of (Z, T) conditionally on $\boldsymbol{X}$. The censoring indicators for Z and T, $\Delta_1 = 1_{\{Z \leqslant C\}}$ and $\Delta = 1_{\{T \leqslant C\}}$, are also observed.

Let $\zeta(t)$ denote the state occupied by the process by time t. It is of our interest to estimate the conditional transition probabilities given $\boldsymbol{X} = \boldsymbol{x}$, $p_{ij}(s, t|\boldsymbol{x}) = P(\zeta(t) = j|\zeta(s) = i, \boldsymbol{X} = \boldsymbol{x})$, $i \leqslant j; i = 1, 2; j = 1, 2, 3$, where s remains fixed and $t > s$. Explicitly, we assume a logit link function so

$$p_{ij}(s, t|\boldsymbol{X}) = \frac{\exp\left(\boldsymbol{X}'\boldsymbol{\beta}_{ij}^{(s)}(t)\right)}{1 + \exp\left(\boldsymbol{X}'\boldsymbol{\beta}_{ij}^{(s)}(t)\right)},$$

where $\boldsymbol{\beta}_{ij}^{(s)}(t)$ is the vector of possibly time-varying coefficients at time t. Other link functions are possible. The model allows for a time-dependent evaluation of the covariate effects on the transition probabilities.

Because of censoring, the individual trajectories may not be completely observed. This implies that the value of $\zeta(t)$ may be unavailable. However, the censored information may be used to introduce suitable estimating equations for $\boldsymbol{\beta}_{ij}^{(s)}(t)$. Let Δ_1^t and Δ^t be the censoring status of Z and T by time t, these are $\Delta_1^t = 1_{\{\min(Z,t) \leqslant C\}}$ and $\Delta^t = 1_{\{\min(T,t) \leqslant C\}}$. To be specific, we consider in the following lines the particular case in which $i = 1$ and $j = 2$. In this case, we have $\{\zeta(t) = j\} = \{\zeta(t) = 2\} = \{Z <= t < T\}$ and therefore

$$\begin{aligned} p_{12}(s, t|\boldsymbol{x}) &= P(Z \leqslant t < T|Z > s, \boldsymbol{X} = \boldsymbol{x}) \\ &= E\left[\frac{1_{\{Z \leqslant t < T\}}\Delta^t}{E(\Delta^t|Z, T, \tilde{Z} > s, \boldsymbol{X} = \boldsymbol{x})}\Big|\tilde{Z} > s, \boldsymbol{X} = \boldsymbol{x}\right] \\ &= E\left[\frac{1_{\{\tilde{Z} \leqslant t < \tilde{T}\}}\Delta^t}{G_{\boldsymbol{x}}^{(s)}(\min(\tilde{T}, t))}\Big|\tilde{Z} > s, \boldsymbol{X} = \boldsymbol{x}\right], \end{aligned} \tag{1}$$

where $G_{\boldsymbol{x}}^{(s)}(t) = P(C \geqslant t|C > s, \boldsymbol{X} = \boldsymbol{x})$ is the conditional survival function of the censoring time given $\boldsymbol{X} = \boldsymbol{x}$ and $C > s$. This equation (1) suggests that the indicator $Y_i(t) = 1_{\{\tilde{Z}_i \leqslant t < \tilde{T}_i\}}$ should be weighted by the random weight $W_{12}^{(s)}(\boldsymbol{X}_i, \tilde{T}_i, \Delta_i^t) = \Delta_i^t G_{\boldsymbol{X}_i}^{(s)}(\min(\tilde{T}_i, t))^{-1}$ in order to eliminate the censoring bias in the evaluation of $p_{12}(s, t|\boldsymbol{x})$. In practice, because $G_{\boldsymbol{x}}^{(s)}$ is unknown, it must be replaced by an estimator $\hat{G}_{\boldsymbol{x}}^{(s)}$, leading to an estimated weight $\hat{W}_{12}^{(s)}$. The estimator $\hat{G}_{\boldsymbol{x}}^{(s)}$ can be constructed by the Kaplan–Meier method if the censoring time is independent of the covariates; otherwise, a semi-parametric estimator based on, for example, Cox regression can be used.

To estimate $\boldsymbol{\beta}_{12}^{(s)}(t)$, we consider the estimating equation

$$\sum_{i:\tilde{Z}_i > s} \frac{\partial p_{12}(s, t|\boldsymbol{X}_i)}{\partial \boldsymbol{\beta}_{12}^{(s)}(t)} \hat{W}_{12}^{(s)}(\boldsymbol{X}_i, \tilde{T}_i, \Delta_i^t)\left[Y_i(t) - p_{12}(s, t|\boldsymbol{X}_i)\right] = 0.$$

As mentioned earlier, s is considered to be fixed and $t \in [a, \tau]$, where τ is the last event time point, that is, $p_{12}(s, a|\boldsymbol{x}) > 0$, and we assume that the survival function of the censoring distribution at τ is larger than zero. Equation (1) indicates that this is an unbiased score function for $\boldsymbol{\beta}_{12}^{(s)}(t)$. In the uncensored case, the solution of this equation is just the ordinary least squares estimator, while in the censored case, it is the minimizer of the sum of squares of differences between the censored binary response $Y_i(t)$ and the predictor $p_{12}(s, t|\boldsymbol{X}_i)$, and the random weights outside the brackets remove the censoring bias. The only terms depending on t in the estimating equation earlier are $Y_i(t)$ and $\hat{W}_{12}^{(s)} = \hat{W}_{12}^{(s)}(\boldsymbol{X}_i, \tilde{T}_i, \Delta_i^t)$, $i = 1, \ldots, n$, which are left unchanged for t between two consecutive points in the set $\{\tilde{Z}_i, \tilde{T}_i : \tilde{Z}_i \leqslant \tilde{T}_i, \tilde{Z}_i > s, \Delta_i = 1\}$.

Therefore, $\hat{\boldsymbol{\beta}}_{12}^{(s)}(t)$ is piecewise constant. Thus, by using standard software for generalized linear models, one can estimate the effect of covariates at each jump point.

Similar issues as those described for $p_{12}(s,t|\boldsymbol{X})$ appear for the other transition probabilities. Of course, the state indicators, the random weights, the set of jump points for the regression parameter, and the subsample to be used in the construction of the score equation will be specific for each particular case. Explicitly, for the transition probabilities $p_{11}(s,t|\boldsymbol{X})$, $p_{13}(s,t|\boldsymbol{X})$, and $p_{23}(s,t|\boldsymbol{X})$, the jump points of $\hat{\boldsymbol{\beta}}_{11}^{(s)}(t)$, $\hat{\boldsymbol{\beta}}_{13}^{(s)}(t)$, and $\hat{\boldsymbol{\beta}}_{23}^{(s)}(t)$ are $\{\tilde{Z}_i : \tilde{Z}_i > s\}$, $\{\tilde{T}_i : \tilde{Z}_i > s\}$, and $\{\tilde{T}_i : \tilde{Z}_i \leqslant s < \tilde{T}_i\}$, respectively. The differences stem from the definition of the binary responses and the random weights, which are $1_{\{\tilde{Z}_i > t\}}$, $1_{\{\tilde{T}_i \leqslant t\}}$, $1_{\{\tilde{T}_i \leqslant t\}}$; and $\Delta_{1i}^t G_{X_i}^{(s)}(\min(\tilde{Z}_i,t))^{-1}$, $\Delta_i^t G_{X_i}^{(s)}(\min(\tilde{T}_i,t))^{-1}$, $\Delta_i^t G_{X_i}^{[s]}(\min(\tilde{T}_i,t))^{-1}$ correspondingly, where $G_{\boldsymbol{x}}^{[s]}(t) = P(C \geqslant t|Z \leqslant s < T, s < C, \boldsymbol{X} = \boldsymbol{x})$ stands for the conditional survival function of C given $\boldsymbol{X} = \boldsymbol{x}$ for the subset of individuals observed in state 2 by time s. To see that the aforementioned weights introduce unbiased score equations, note that we have

$$\begin{aligned}
p_{11}(s,t|\boldsymbol{x}) &= P(Z > t|Z > s, \boldsymbol{X} = \boldsymbol{x}) \\
&= E\left[\frac{1_{\{Z>t\}}\Delta_1^t}{E(\Delta_1^t|Z,\tilde{Z} > s, \boldsymbol{X} = \boldsymbol{x})}\Big|\tilde{Z} > s, \boldsymbol{X} = \boldsymbol{x}\right] \\
&= E\left[\frac{1_{\{\tilde{Z}>t\}}\Delta_1^t}{G_{\boldsymbol{x}}^{(s)}(\min(\tilde{Z},t))}\Big|\tilde{Z} > s, \boldsymbol{X} = \boldsymbol{x}\right], \\
p_{13}(s,t|\boldsymbol{x}) &= P(T \leqslant t|Z > s, \boldsymbol{X} = \boldsymbol{x}) \\
&= E\left[\frac{1_{\{T\leqslant t\}}\Delta^t}{E(\Delta^t|Z,T,\tilde{Z} > s, \boldsymbol{X} = \boldsymbol{x})}\Big|\tilde{Z} > s, \boldsymbol{X} = \boldsymbol{x}\right] \\
&= E\left[\frac{1_{\{\tilde{T}\leqslant t\}}\Delta^t}{G_{\boldsymbol{x}}^{(s)}(\min(\tilde{T},t))}\Big|\tilde{Z} > s, \boldsymbol{X} = \boldsymbol{x}\right],
\end{aligned}$$

and

$$\begin{aligned}
p_{23}(s,t|\boldsymbol{x}) &= P(T \leqslant t|Z \leqslant s < T, \boldsymbol{X} = \boldsymbol{x}) \\
&= E\left[\frac{1_{\{T\leqslant t\}}\Delta^t}{E(\Delta^t|Z,T,\tilde{Z} \leqslant s < \tilde{T}, \boldsymbol{X} = \boldsymbol{x})}\Big|\tilde{Z} \leqslant s < \tilde{T}, \boldsymbol{X} = \boldsymbol{x}\right] \\
&= E\left[\frac{1_{\{\tilde{T}\leqslant t\}}\Delta^t}{G_{\boldsymbol{x}}^{[s]}(\min(\tilde{T},t))}\Big|\tilde{Z} \leqslant s < \tilde{T}, \boldsymbol{X} = \boldsymbol{x}\right].
\end{aligned}$$

As discussed earlier for $G_{\boldsymbol{X}}^{(s)}$, $G_{\boldsymbol{X}}^{[s]}$ can be estimated using the Kaplan–Meier estimator or a semi-parametric regression approach (e.g., Cox model) by considering the corresponding subsample $\{i : \tilde{Z}_i \leqslant s < \tilde{T}_i\}$ for the last case. Here, again, the weights are time dependent, but between two consecutive jump points remain constant; hence, the stepwise nature of $\hat{\boldsymbol{\beta}}_{ij}^{(s)}(t)$ remains. These weights, as defined, are non-zero for large censored observations, which is convenient in order to control the variance at the right tail of the survival time distribution.

Asymptotic properties of the introduced estimators under regularity conditions can be established as usual in M-estimation. For example, if the number of time points in the score equation is fixed, the proposed estimator is the solution of a weighted version of the standard generalized linear model score equation. Therefore, if the weights were known, the uniqueness and consistency of the estimator would hold, provided that the first derivative of the score equation is continuous and invertible in $\boldsymbol{\beta}_{ij}^{(s)}(t)$ and converges uniformly in an open-neighborhood of $\boldsymbol{\beta}_{ij}^{(s)}(t)$ [14]. Note that the uniform convergence of the derivative of the score equation follows for sufficiently smooth link functions. To deal with the random weights, assume that both $\hat{G}_{\boldsymbol{x}}^{(s)}(.)$ and $\hat{G}_{\boldsymbol{x}}^{[s]}(.)$ converge uniformly to their respective limits and that the survival function of the censoring distribution is bounded away from zero on $[0,\tau]$. Then, under such conditions, it is easy to see that the consistency result holds for the random weights too, in the finite-t case. When we consider the score equations jointly for all $t \in [s,\tau]$, a uniform consistency result can still be derived from [14] if the first derivative of the score equation converges to an invertible function uniformly on t (e.g., [15]). To fulfill this condition, it suffices to impose the boundedness of all the arguments in the score equation, noting again the smoothness of our link function.

On the other hand, the asymptotic normality can be obtained from the decomposition of $\sqrt{n}\big(\hat{\boldsymbol{\beta}}_{ij}^{(s)}(t) - \boldsymbol{\beta}_{ij}^{(s)}(t)\big)$ as a sum of centered iid random variables plus a negligible remainder, by applying the Central Limit Theorem (CLT) (see also Liang and Zeger [16] for asymptotic results in generalized estimating equations). To be more specific, for $p_{12}(s,t|\boldsymbol{x})$, we have the following iid decomposition:

$$\sqrt{n}\{\underline{I}_{12}(t,\boldsymbol{\beta}_{12}(t))\}^{-1}\sum_{i:\tilde{Z}_i>s}\underline{\mathrm{W}}_{12i}(t)$$

where

$$\underline{\mathrm{I}}_{12}(t,\boldsymbol{\beta}_{12}(t)) = E\left(\left[\frac{\partial p_{12}(s,t|\boldsymbol{X})}{\partial\boldsymbol{\beta}_{12}^{(s)}(t)}\right]^T W_{12}^{(s)}(\boldsymbol{X}_i,\tilde{T}_i,\Delta_i^t)\left[\frac{\partial p_{12}(s,t|\boldsymbol{X})}{\partial\boldsymbol{\beta}_{12}^{(s)}(t)}\right]\right)$$

and

$$\underline{\mathrm{W}}_{12i}(t) = \left[\frac{\partial p_{12}(s,t|\boldsymbol{X}_i)}{\partial\boldsymbol{\beta}_{12}^{(s)}(t)}W_{12}^{(s)}(\boldsymbol{X}_i,\tilde{T}_i,\Delta_i^t)[Y_i(t)-p_{12}(s,t|\boldsymbol{X}_i)]+\psi_{12i}(t)\right].$$

The random weights are responsible for the ψ_{12} terms, whose explicit form is given in the Appendix. Interestingly, their contribution is typically very small and ignoring them will give conservative standard errors as argued in [8]. For $p_{23}(s,t|\boldsymbol{X})$, the iid decomposition is as follows:

$$\sqrt{n}\{\underline{I}_{23}(t,\boldsymbol{\beta}_{23})\}^{-1}\sum_{i:\tilde{Z}_i\leqslant s<\tilde{T}_i}\underline{\mathrm{W}}_{23i}(t)$$

where

$$\underline{\mathrm{I}}_{23}(t,\boldsymbol{\beta}_{23}(t)) = E\left(\left[\frac{\partial p_{23}(s,t|\boldsymbol{X})}{\partial\boldsymbol{\beta}_{23}^{(s)}(t)}\right]^T \frac{\Delta^t}{G_{\boldsymbol{x}}^{[s]}(\min(\tilde{T},t))}\left[\frac{\partial p_{23}(s,t|\boldsymbol{X})}{\partial\boldsymbol{\beta}_{23}^{(s)}(t)}\right]\right)$$

and

$$\underline{\mathrm{W}}_{23i}(t) = \left[\frac{\partial p_{23}(s,t|\boldsymbol{X}_i)}{\partial\boldsymbol{\beta}_{23}^{(s)}(t)}.\frac{\Delta^t}{G_{\boldsymbol{x}}^{[s]}(\min(\tilde{T}_i,t))}[1_{\{\tilde{T}_i\leqslant t\}}-p_{23}(s,t|\boldsymbol{X}_i)]+\psi_{23i}(t)\right],$$

where again the explicit form of the ψ_{23} is given in the Appendix. The same argument can be used to prove the asymptotic normality for the other transition probabilities.

3. Simulation study

We investigate the performance of our method through simulations. A binary covariate X is considered, $X \sim Ber(p), p = 1/2$. The progressive illness–death model is simulated by means of three latent transition times: T_{12}, T_{13}, and T_{23}, where T_{ij} denotes the potential transition time from state i to state j. Given X, the transition times T_{12} and T_{13} are simulated as independent random variables as follows: $T_{12}|X = x$; $T_{13}|X = x$ are generated from an exponential model with rates 1 and 0.1; and 1.5, 0.1, and $T_{23}|T_{12} = u, X = x$ is generated from exponential model with rates $u * 0.7$ and $u * 0.3$ for $x = 0$ and $x = 1$, respectively; here the dependency of T_{23} on T_{12} violates the Markov condition. The variables $Z = \min(T_{12}, T_{13})$, $\rho = 1_{\{T_{12}\leqslant T_{13}\}}$, $T = Z + \rho T_{23}$ are then computed. An independent censoring time C is generated from a uniform model with minimum value 0 and maximum values 64 and 11, which correspond to 15% and 45% of censoring on the total survival time T, respectively.

The observed variable is finally given by $(X, \tilde{Z}, \tilde{T})$ where $\tilde{Z} = \min(Z, C)$, $\tilde{T} = \min(T, C)$.

When the link function is logistic, we have

$$p_{ij}(s,t|X=x) = \frac{\exp\left(\boldsymbol{X}'\boldsymbol{\gamma}_{ij}(s,t)\right)}{1+\exp\left(\boldsymbol{X}'\boldsymbol{\gamma}_{ij}(s,t)\right)} = \frac{\exp\left(\beta_{ij_1}(s,t)(1-x)+\beta_{ij_2}(s,t)x\right)}{1+\exp\left(\beta_{ij_1}(s,t)(1-x)+\beta_{ij_2}(s,t)x\right)}$$

where $i = 1, 2; j = 1, 2, 3$, $\boldsymbol{\gamma}'_{ij}(s,t) = (\beta_{ij_1}(s,t), \beta_{ij_2}(s,t) - \beta_{ij_1}(s,t))$ and $\boldsymbol{X}' = (1, X)$. Thus, the simulated $\beta_{ij_1}(s,t)$ and $\beta_{ij_2}(s,t)$ functions in the logistic model are given by

$$\beta_{ij_1}(s,t) = log\frac{p_{ij}(s,t|x=0)}{1-p_{ij}(s,t|x=0)} \quad , \quad \beta_{ij_2}(s,t) = log\frac{p_{ij}(s,t|x=1)}{1-p_{ij}(s,t|x=1)}$$

whose expressions may be directly obtained from the conditional distributions of the T_{ij}'s. The effect of a covariate increase (from $x = 0$ to $x = 1$) on $p_{ij}(s,t|X = x)$ is controlled by $\beta_{ij}(s,t) = \beta_{ij_2}(s,t) - \beta_{ij_1}(s,t)$.

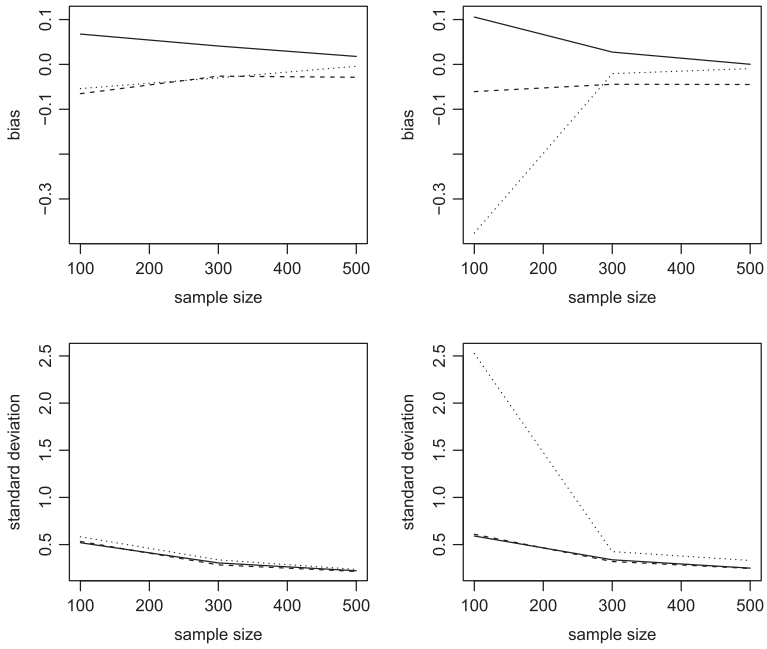

Figure 2. Bias (top) and standard deviation (bottom) of $\hat{\boldsymbol{\beta}}_{ij}$ for $(i,j) = (1,2)$ (solid lines), $(i,j) = (1,3)$ (dashed lines), and $(i,j) = (2,3)$ (dotted lines). Censoring percentage 15% (left) and 45% (right).

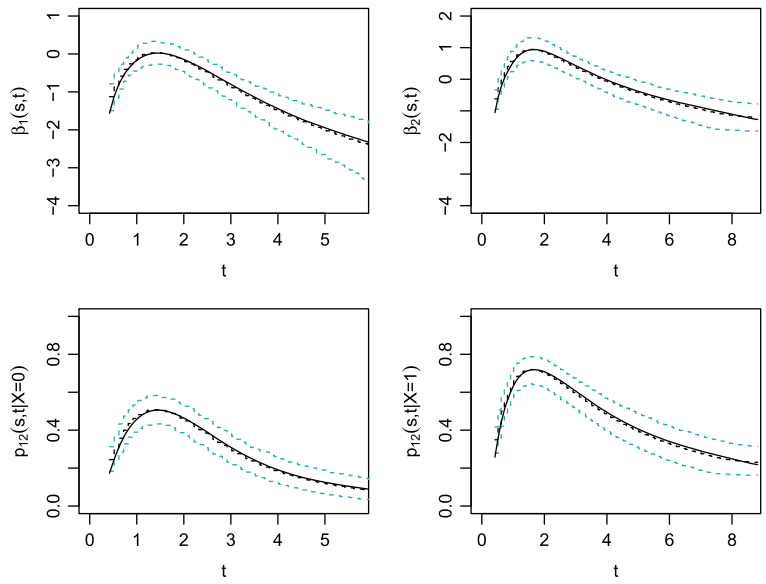

Figure 3. Estimated regression coefficients and conditional transition probabilities $p_{12}(s,t|X)$ averaged along 1000 Monte Carlo trials, together with 95% oscillation limits: $s = 0.22$, $n = 500$, and censoring percentage 45%. The dotted line equals the estimator, whereas the solid line is the true value. [Colour figure can be viewed at wileyonlinelibrary.com]

In the simulation, 1000 samples with sample sizes $n = 100$, $n = 300$, and $n = 500$ are generated.

In Figure 2, the bias and the standard deviation of the estimated effects ($\hat{\beta}_{ij}$) on transition probabilities p_{ij}, $i = 1, 2, j = 2, 3$, $i < j$ averaged along the 1000 trials are given. For this, we take specific values for the pair of time points (s, t), $s < t$. For $p_{12}(s, t)$ and $p_{13}(s, t)$, $s = 0.22$ and $t = 2.82$ equal the 25% quantile of Z, and the 50% quantiles of T given $Z > s$, respectively, while for $p_{23}(s, t|X = x)$, we consider $s = 1.10$ and $t = 6.10$ corresponding to the 75% quantile of Z and 50% quantile of T given $Z \leqslant s < T$.

In Figure 2, standard deviation increases with an increasing censoring degree and decreases with an increasing sample size. Both features were expected. The absolute bias is in general of smaller order compared with the standard deviation. Results for other pairs (s, t) (not shown) were similar to those in Figure 2.

In Figure 3 we report, for a fixed s, the estimators $\hat{\beta}_1(s, t)$ and $\hat{\beta}_2(s, t)$ for the transition probability p_{12} along time t averaged along the 1000 Monte Carlo paths, for the case $n = 500$ and censoring degree 45%; the averages of the resulting conditional transition probabilities given $X = 0$ and $X = 1$ are also reported. In this figure, the true curves being estimated and the 95% pointwise oscillation limits of the estimators are included too. The lower and upper oscillation limits are, respectively, 0.025 and 0.975 quantiles of the estimates along 1000 Monte Carlo trails. The dashed line equals the estimator, whereas the straight line is the true value. We take $s = 0.22$, and the time endpoints are taken as the 95% quantile of total survival time, $\tilde{T}$, given $\tilde{Z} > s$ of the subgroup $X = 0$ and $X = 1$, respectively, for 45% censoring degree, that is, $\tau = 5.70$ and $\tau = 8.57$.

In Figures 3, we see that the estimators accurately estimate their respective targets. Also, the 95% oscillation intervals get wider as time increases. Note that the endpoint for the subgroup $X = 0$ is smaller than the one for $X = 1$, because the 95% quantile of $\tilde{T}$ is smaller for $X = 0$. Then, the plots inform on the performance of the estimators for the time-varying coefficients on a time interval, which skips the upper 5% tail of $\tilde{T}$ in each subgroup; this makes the plots for β_1 and β_2 ,respectively, comparable. We have constructed the plots in Figure 3 for the cases $n = 100$ and $n = 300$ too (not shown). By comparing these plots to those corresponding to $n = 500$, one sees that the accuracy of the estimators greatly improves as the sample size grows, as expected. The similar results have been concluded from the plots corresponding to the transition probabilities p_{11}, p_{13}, p_{23} (not shown again).

4. Lupus data

Systemic lupus erythematosus (SLE), often abbreviated as lupus, is a systemic autoimmune disease (or autoimmune connective tissue disease) in which the body's immune system mistakenly attacks healthy tissue. SLE, the most common and severe type of lupus, affects many internal organs in the body. SLE most often harms the heart, joints, skin, lungs, blood vessels, liver, kidneys, and nervous system. Because the information about the clinical characteristics of SLE usually comes from small number of patients, the Spanish Society of Rheumatology (SER) promoted the creation of a large multicenter registry of SLE patients, aimed to increase the overall knowledge of the disease. RELESSER was conducted by members of the Systematic Autoimmune Disease Study Group of the SER, and it involved 45 centers homogeneously spread across Spain [17].

In this section, we study the lupus (SLE) dataset of the RELESSER registry. Because damage is associated with mortality in SLE patients (see Pego-Reigosa *et al.* [18] and references therein), we consider it as a relevant intermediate event in our analysis. We model the lupus dataset by the progressive illness–death model in Figure 4, where the 'Healthy' state stands for no damage, and the 'Diseased' state refers to the

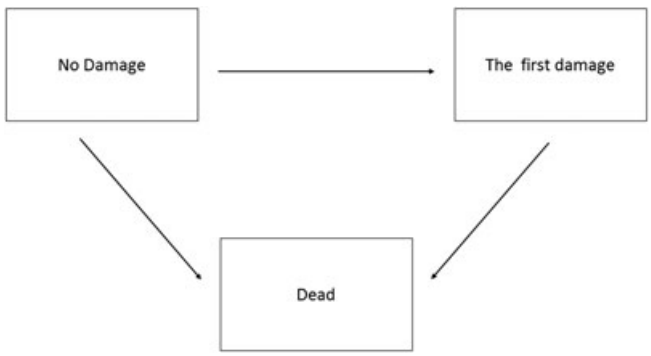

Figure 4. Progressive illness–death model for the lupus data.

Table I. Lupus data. Number of patients at risk, damaged, dead, and censored.

	At risk	Damage	Death	Censored
Age at diagnosis				
Young (< 50 years)	2913	1063	126	2787
Elderly(≥ 50 years)	547	276	73	474
Missing	112			
Sex				
Male (0)	333	160	30	303
Female (1)	3130	1180	169	2961
Missing	109			
Ethnicity				
Caucasian (0)	3284	1289	191	3093
Hispanic (1)	183	52	8	175
Missing	105			

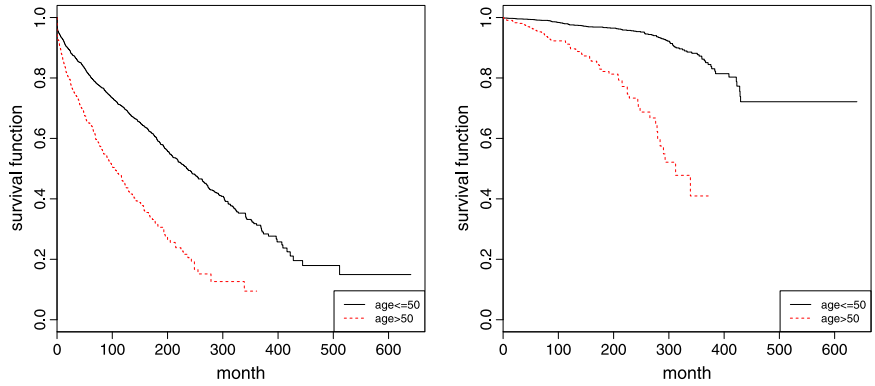

Figure 5. Lupus data. Damage-free survival (left) and total survival (right) for two age groups: young (< 50 years) and elderly(≥ 50 years). [Colour figure can be viewed at wileyonlinelibrary.com]

first among the following damages provoked by lupus: ocular, neuropsychiatric, pulmonary, cardiovascular, peripheral vascular, gastrointestinal, musculoskeletal , skin, diabetes, malignancy, and premature gonadal failure. Specifically, we analyze the $n = 3572$ cases in RELESSER corresponding to the patients for which the (maybe censored) transition times were available. The number of patients at risk, damaged, dead, and censored according to the epidemiological variables age, sex, and ethnicity for this dataset are described in Table I. We restricted our attention to Caucasian and Hispanic patients because other ethnicity groups were heavily censored and reported very small sample sizes.

4.1. Preliminary survival analysis

Damage-free survival and total survival curves for each group of age, sex, and ethnicity were obtained by a direct application of the Kaplan–Meier estimator to the (censored) damage-free survival times and total survival times. These curves showed that age at diagnosis and sex are important for prognosis (with a poorer survival for older patients and males), while the effect of ethnicity was less clear (crossing curves). In Figure 5, we report the survival curves corresponding to age for illustration. The impact of the epidemiological variables on the damage-free survival and total survival was assessed through multivariate Cox proportional hazards regression too. Cox regression reported a significant effect of age and sex (p-values $p_{age} = (p < 0.001, p < 0.001)$ and $p_{sex} = (p < 0.001, 0.0315)$ for the damage-free and total survival, respectively), while the effect of ethnicity was significant on total survival ($p_{eth} = 0.0352$, the relative hazard Hispanic vs. Caucasian was 2.17), but it was not significant for damage-free survival ($p_{eth} = 0.2090$). In subsection 4.2, we apply the time-varying coefficient approach proposed in this paper to investigate the adjusted effects of the epidemiological covariates, which results in a more flexible multivariate analysis. Besides, a dynamic viewpoint is included by estimating covariate effects on transition probabilities $p_{ij}(s, t)$ for various values of s.

4.2. *Direct modeling approach*

In this subsection, we analyze the lupus data by employing the time-varying coefficient approach introduced in Section 2. Results corresponding to estimated (time-varying) effects adjusted for the epidemiological covariates age, sex, and ethnicity, together with 95% pointwise bootstrap confidence bands based on nonparametric resampling and normal method (normal approximation of two-sided nonparametric confidence interval), are displayed in Figure 6. The first two rows in Figure 6 correspond to $p_{11}(s,t)$ for $s = 0$ (top) and $s = 24$ (middle), where time is in months. That is, covariate effects on damage-free both from the date of diagnosis ($s = 0$) and 2 years later ($s = 24$) are the focus. Bottom row in Figure 6 depicts the effect of covariates on $p_{12}(0,t)$. For $p_{2j}(s,t)$, $j = 2,3$, no covariate was significant for the considered values of s (to save space, results are not shown).

From Figure 6(a), we can see that the adjusted effect of age on $p_{11}(0,t)$ is negative and time-varying, increasing in absolute value along time. That is, the damage-free survival decreases with age, the effect being stronger for older patients. On the other hand, Figure 6(d) indicates that this effect becomes roughly constant when the analysis is restricted to the patients without damage 2 years after diagnosis. Besides, the adjusted effect of age on $p_{12}(0,t)$ is positive at early times but negative at late times (Figure 6(g)). A possible explanation for this is that older patients have more chances to develop damages, but a larger probability of dying too. Indeed, the effect of age on the cumulative probability of death, $p_{13}(0,t)$, is positive and increasing along time (not shown).

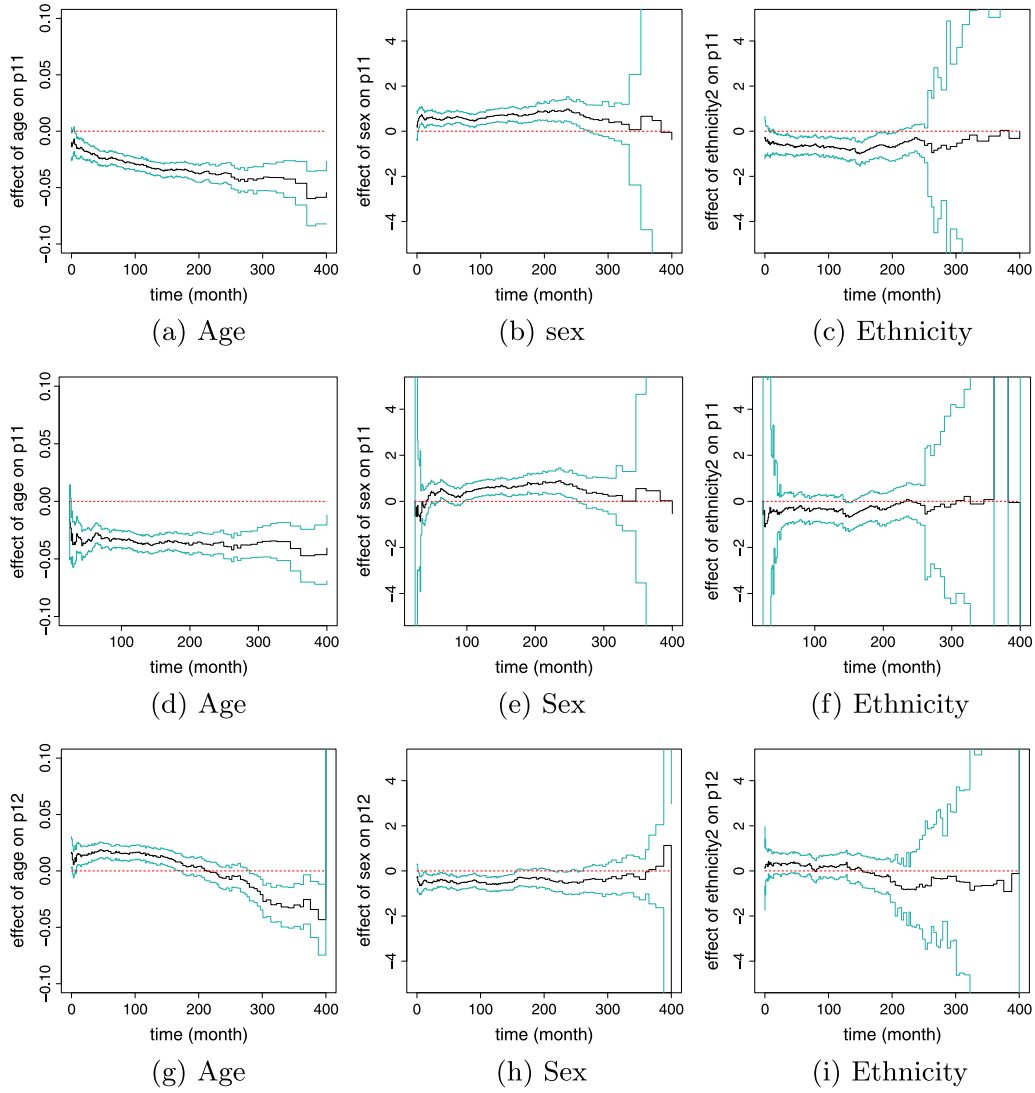

Figure 6. Lupus data. Estimated adjusted effect of age, sex, and ethnicity (from left to right) on $p_{11}(s = 0,t)$ (top), $p_{11}(s = 24,t)$ (middle), and $p_{12}(0,t)$ (bottom), with 95% bootstrap pointwise confidence limits. [Colour figure can be viewed at wileyonlinelibrary.com]

As for the effect of sex, from Figure 6(b), it is seen that females have a larger damage-free survival, and that the adjusted effect of sex on $p_{11}(0,t)$ is constant along time. However, for the patients without damage 2 years after diagnosis, no significant differences are found between male and female damage-free survival functions in the following 8 years (Figure 6(e)), possibly because of the decreased sample size. The plot corresponding to $p_{12}(0,t)$ (Figure 6(h)) indicates that the larger damage-free survival of the females could be due to a relatively smaller probability of visiting the intermediate state (damage), rather than to a smaller probability of death. This was confirmed by computing the adjusted effects of sex on $p_{13}(0,t)$ (not shown), which did not reach significance.

From Figure 6(c), we see that, when adjusting for age and sex, Hispanic patients have a damage-free survival significantly smaller than that of Caucasian; their total survival is smaller too (effects not shown). These are interesting findings compared with the preliminary survival analysis earlier, which was somehow inconclusive with respect to ethnicity. Unlike for age and sex, ethnic groups have crossing Kaplan–Meier curves and, therefore, proportional hazards assumption may fail. The adjusted effect of ethnicity is constant along time in both cases (damage-free survival, total survival). However, the effect vanishes when analyzing alive and damage-free patients two years after lupus diagnosis (Figure 6(f)). On the other hand, no significant effect of ethnicity was found for $p_{12}(0,t)$ (Figure 6(i)) meaning that, for $s=0$, Hispanic patients worsen their damage-free survival (compared with Caucasian) by increasing the probability of death.

The pointwise confidence limits in Figure 6 are occasionally wide for early and late times. Indeed, the estimator for the time-varying coefficient is not accurate at time points where the state occupancy indicators are strongly unbalanced or heavily censored. In practice, this means that significance will only be reached along some compact time interval where there exists enough, well-balanced sampling information.

In Figure 7, we depict the conditional transition probabilities for the patients with and without damage 1 year after diagnosis. Conditional curves corresponding to Caucasian males and females are separately displayed, while covariate age is fixed at the average. In particular, it is seen how $p_{11}(12,t)$ (respectively $p_{12}(12,t)$) is larger (resp. smaller) for the Caucasian females when adjusting for age, the differences being less clear for the other transition probabilities. The provided results were obtained by using the logit link

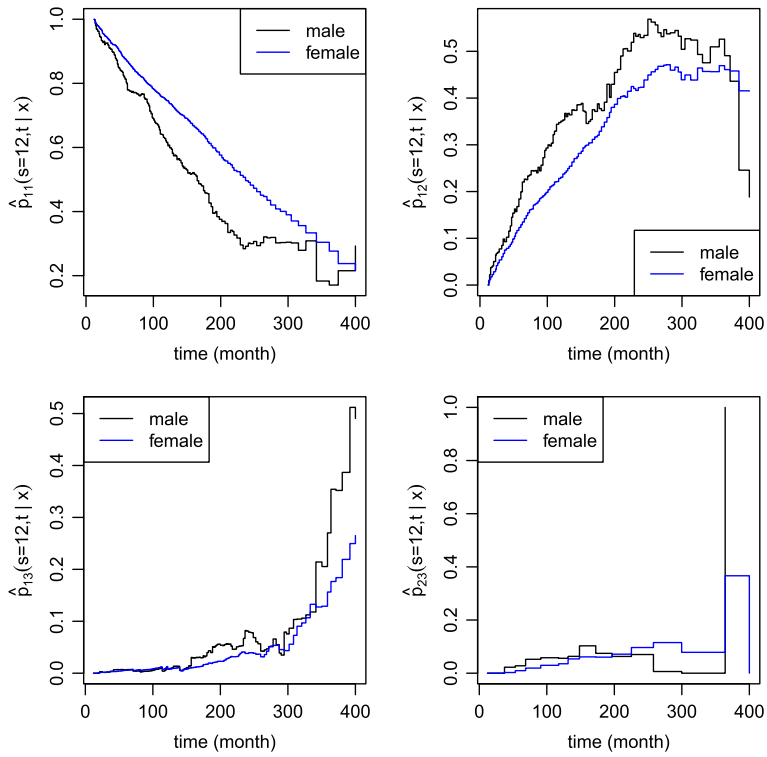

Figure 7. Lupus data. Adjusted transition probabilities $p_{11}(12,t|\boldsymbol{x})$, $p_{12}(12,t|\boldsymbol{x})$, $p_{13}(12,t|\boldsymbol{x})$, and $p_{23}(12,t|\boldsymbol{x})$ for two different groups of sex (male and female) based on logit link function, when evaluated at the average of age and for Caucasian ethnicity.

function. Adjusted transition probabilities based on the probit link were also computed, giving almost identical results to those in Figure 7 (not shown). Then, we can sum up that the choice of the link function has a minor impact in the resulting estimator. This robustness property of the method is interesting, being due to the fact that the proposed approach imposes no parametric structure along time.

5. Discussion

In this paper, a new method to incorporate covariates in the transition probability matrix for the right-censored progressive illness–death model has been proposed. The new method employs direct binomial regression in the sense discussed by Scheike *et al.* [8] for competing risks. The flexibility of the method, which imposes no structure for the covariate effects along time, is one of its main advantages. This may prevent the researcher from ignoring covariate effects that could remain undetected in a less flexible modeling approach. The proposed method is able to deal with multiple covariates and non-Markov structures. This entails further flexibility. The lupus dataset analyzed in Section 4 could violate the Markov condition if, for the patients with damage at a given time point, the date of first damage was associated to the survival time. Preliminary analysis of such possible association through a proportional hazards model, including the epidemiological variables and the time to first damage as covariates, reported no significant deviation from Markovianity ($p = 0.4$). In any case, the time-varying coefficient approach in this paper is robust to this regard, which can be considered as an important property of the method.

This piece of work opens new interesting research lines for the future. For example, the general idea behind the construction of the proposed estimator can be used in principle to introduce covariate effects for transition probabilities in progressive multi-state models other than the illness–death model. Scheike and Zhang [19] considered this problem in the case of the so-called occupation probabilities in a fairly general multi-state process. They also pointed out some drawbacks of our estimating approach, such as its possible inefficiency and violation of natural constraints; still, the flexibility and relative good behavior of the direct binomial regression approach makes the method recommendable. Another possible extension of the estimator proposed in this paper is to incorporate left-truncation; in the truncated setting, the random weights must be corrected to properly compensate for the observational bias. This issue is currently under investigation.

Appendix

Here, we present the ψ terms we mentioned in Section 2. First, we follow the martingale representation of the Kaplan–Meier estimator (as in [9]) and, for simplicity, we suppose that the censoring is independent of the covariates; then, our notation for $G_{\boldsymbol{x}}^{(s)}$ and $\hat{G}_{\boldsymbol{x}}^{(s)}$ reduces to $G^{(s)}$ and $\hat{G}^{(s)}$, respectively. Write

$$\begin{aligned}
\hat{W}_{12}^{(s)}(\boldsymbol{X}_i, \tilde{T}_i, \Delta_i^t) - W_{12}^{(s)}(\boldsymbol{X}_i, \tilde{T}_i, \Delta_i^t) &= \Delta_i^t \left(\frac{1}{\hat{G}^{(s)}(\min(\tilde{T}_i, t)^-)} - \frac{1}{G^{(s)}(\min(\tilde{T}_i, t)^-)} \right) \\
&= \frac{\Delta_i^t}{G^{(s)}(\min(\tilde{T}_i, t)^-)} \left(\frac{G^{(s)}(\min(\tilde{T}_i, t))}{\hat{G}^{(s)}(\min(\tilde{T}_i, t)^-)} - 1 \right) \\
&= \frac{\Delta_i^t}{G^{(s)}(\min(\tilde{T}_i, t)^-)} \sum_{j:\tilde{Z}_j > s} \int_0^\tau 1_{\{s \leqslant \min(\tilde{T}_i, t)\}} \frac{1}{p_1(u)} dM_{1j}^c(u) \\
&\quad + o_p(n^{-1/2})
\end{aligned}$$

where $p_1(t) = \sum_{i:\tilde{Z}_i > s} 1_{\{\tilde{T}_i \geqslant t\}}$ and M_{1i}^c's are the basic censoring time martingales of the form

$$\begin{aligned}
\hat{M}_{1i}^c(t) &= N_i^c(t) - \int_0^t 1_{\{\tilde{T}_i \geqslant u\}} d\hat{\Lambda}_1^c(u) \\
\hat{\Lambda}_1^c(t) &= \sum_{i:\tilde{Z}_i > s} \int_0^t \frac{1_{\{p_1(u) > 0\}}}{p_1(u)} dN_i^c(u).
\end{aligned}$$

where $N_i^c(t) = 1_{\{\tilde{T}_i \leqslant t, \tilde{T}_i = C_i\}}$.

As a consequence, we have

$$\sum_{i:\tilde{Z}_i > s} \frac{\partial p_{12}(s,t|\boldsymbol{X}_i)}{\partial \boldsymbol{\beta}_{12}^{(s)}(t)} \left[Y_i(t) - p_{12}(s,t|\boldsymbol{X}_i)\right] (\hat{W}_{12}^{(s)}(\boldsymbol{X}_i, \tilde{T}_i, \Delta_i^t) - W_{12}^{(s)}(\boldsymbol{X}_i, \tilde{T}_i, \Delta_i^t)) = \sum_i \psi_{12i}(t) + o_p(n^{-1/2})$$

where

$$\psi_{12i}(t) = \int_0^\tau \frac{q_{12}(u,t)}{\pi_1(u)} dM_{1i}^c(u)$$
$$\pi_1(u) = \lim_p p_1(u)/n$$
$$q_{12}(u,t) = \lim_p Q_{12}(u,t)/n$$
$$Q_{12}(u,t) = \sum_{j:\tilde{Z}_j > s} \Delta_j^t \frac{\partial p_{12}(s,t|\boldsymbol{X}_i)}{\partial \boldsymbol{\beta}_{12}^{(s)}(t)} \left[Y_i(t) - p_{12}(s,t|\boldsymbol{X}_i)\right] \frac{1_{\{s \leqslant \min(\tilde{T}_i, t)^-\}}}{G^{(s)}(\min(\tilde{T}_i, t)^-)}.$$

Similarly , for the ψ terms appearing in the decomposition pertaining to p_{23}, we have

$$\psi_{23i}(t) = \int_0^\tau \frac{q_{23}(u,t)}{\pi_2(u)} dM_{2i}^c(u)$$
$$q_{23}(u,t) = \lim_p Q_{23}(u,t)/n$$
$$Q_{23}(u,t) = \sum_{j:\tilde{Z}_j \leqslant s < \tilde{T}_j} \Delta_j^t \frac{\partial p_{23}(s,t|\boldsymbol{X}_j)}{\partial \boldsymbol{\beta}_{23}^{(s)}(t)} \left[1_{\{\tilde{T}_j \leqslant t\}} - p_{23}(s,t|\boldsymbol{X}_j)\right] \frac{1_{\{s \leqslant \min(\tilde{T}_j, t)\}}}{G^{[s]}(\min(\tilde{T}_j, t)^-)},$$

$$\pi_2(u) = \lim_p p_2(u)/n$$
$$p_2(u) = \sum_{i:\tilde{Z}_i \leqslant s < \tilde{T}_i} 1_{\{\tilde{T}_i \geqslant t\}}$$

and

$$\hat{M}_{2i}^c(t) = N_i^c(t) - \int_0^t 1_{\{\tilde{T}_i \geqslant u\}} d\hat{\Lambda}_2^c(u)$$
$$\hat{\Lambda}_2^c(t) = \sum_{i:\tilde{Z}_i \leqslant s < \tilde{T}_i} \int_0^t \frac{1_{\{p_2(u)>0\}}}{p_2(u)} dN_i^c(u).$$

Acknowledgements

We thank an anonymous reviewer and the associate editor for comments and suggestions that have improved the paper. We especially thank Rheumatologists José-María Pego-Reigosa and Íñigo Rúa-Figueroa for providing the lupus dataset and for an enlightening discussion and interpretation of the results. Thanks also to Vanesa Balboa for cleaning and preparing the lupus dataset from RELESSER registry. This work was supported by funding from the European Community's Seventh Framework Programme FP7/2011: Marie Curie Initial Training Network MEDIASRES ("Novel Statistical Methodology for Diagnostic/Prognostic and Therapeutic Studies and Systematic Reviews"; www.mediasres-itn.eu) with the Grant Agreement Number 290025. Third author was supported by the Grant MTM2014-55966-P of the Spanish Ministerio de Economía y Competitividad and by CINBIO - Centro Singular de Investigación de Galicia 2016–2019, Consellería de Cultura, Educación e Ordenación Universitaria (FEDER support included).

References

1. Hougaard P. *Analysis of Multivariate Survival Data*. Springer: New York, 2000.
2. Aalen O, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* 1978; **5**:141–150.
3. Meira-Machado L, de Uña-Álvarez J, Cadarso- Suárez C. Non-parametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis* 2006; **12**:325–344.

4. Allignol A, Beyersmann J, Gerds T, Latouche A. A competing risks approach for nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis* 2013; **20**:495–513.
5. Titman A. Transition probability estimates for non-Markov multistate models. *Biometrics* 2015; **71**:1034–1041.
6. de Uña-Álvarez J, Meira-Machado L. Nonparametric estimation of transition probabilities in the non-Markov illness-death model: a comparative study. *Biometrics* 2015; **71**:364–375.
7. Aalen O, Borgan O, Fekjaer H. Covariate adjustment of event histories estimated from Markov chains: the additive approach. *Biometrics* 2001; **57**:993–1001.
8. Scheike TH, Zhang MJ, Gerds TA. Predicting cumulative incidence probability by direct binomial regression. *Biometrika* 2008; **95**:205–220.
9. Fine JP, Gray RJ. A proportional hazards model for subdistribution of a competing risk. *Journal of the American Statistical Association* 1999; **94**:496–509.
10. Andersen PL, Klein JP, Rosthoj S. 2003 Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* 2003; **90**:15–27.
11. Klein JP, Andersen PK. Regression modeling of competing risks data based on pseudo values of the cumulative incidence function. *Biometrics* 2005; **61**:223–229.
12. Gerds TA, Scheike TH, Andersen PK. Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in Medicine* 2012; **31**:3921–3930.
13. Meira-Machado L, de Uña-Álvarez J, Datta S. Nonparametric estimation of conditional transition probabilities in a non-Markov illness-death model. *Computational Statistics* 2015; **30**:377–397.
14. Foutz RV. On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association* 1977; **72**:147–148.
15. Fine JP, Yan J, Kosorok MR. Temporal process regression. *Biometrika* 2004; **91**:683–703.
16. Liang K, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
17. Rúa-Figueroa I, López-Longo FJ, Calvo-Alén J, Galindo-Izquierdo M, Loza E, García de Yebenes MJ, Pego-Reigosa JM, Grupo de trabajo en Enfermedades Autoinmunes Sistémicas de la Sociedad Española de Reumatología (EAS-SER)de la Unidad de Investigación de la Sociedad Espanola de Reumatología (UI-SER) de la Unidad de Investigación de la Sociedad Espanola de Reumatología (UI-SER). National registry of patients with systemic lupus erythematosus of the Spanish Society of Rheumatology: objectives and methodology. *Reumatol Clin* 2014; **10**:17–24.
18. Pego-Reigosa JM, Lois-Iglesias A, Rúa-Figueroa I, Galindo M, Calvo-Alén J, de Uña-Álvarez J, Balboa Barreiro V, Ruan JI, Olivé A, Rodríguez-Gómez M, Nebro AF, Andrés M, Erausquin C, Tomero E, Rubio LH, Isacelaya EU, Freire M, Montilla C, Sánchez-Atrio AI, Santos-Soler G, Zea A, Díez E, Narváez J, Blanco-Alonso R, Silva-Fernández L, Ruiz-Lucea ME, Fernández-Castro M, Hernández-Beriain JA, Gantes-Mora M, Hernández-Cruz B, Pérez-Venegas J, Pecondón-Español A, Fernández-Cid CM, Ibáñez-Barcelo M, Bonilla G, Torrente-Segarra V, Castellví I, Alegre JJ, Calvet J, de la Fuente JLM, Raya E, Vázquez-Rodríguez TR, Quevedo-Vila V, Muñoz-Fernández S, Otón T, Rahman A, López-Longo FJ. Relationship between damage clustering and mortality in systemic lupus erythematosus in early and late stages of the disease: cluster analyses in a large cohort from the Spanish Society of Rheumatology Lupus Registry. *Rheumatology* 2016; **55**:1243–1250.
19. Scheike TH, Zhang M-J. Direct modelling of regression effects for transition probabilities in multistate models. *Scandinavian Journal of Statistics* 2007; **34**:17–32.