Universidade de Vigo

Conditional Transition Probabilities in a non-Markov Illness-death Model

Luís Meira-Machado, Jacobo de Uña-Álvarez and Somnath Datta

Report 12/05

Discussion Papers in Statistics and Operation Research

Departamento de Estatística e Investigación Operativa

Facultade de Ciencias Económicas e Empresariales
Lagoas-Marcosende, s/n · 36310 Vigo
Tfno.: +34 986 812440 - Fax: +34 986 812401
http://webs.uvigo.es/depc05/
E-mail: depc05@uvigo.es

Universidade de Vigo

Conditional Transition Probabilities in a non-Markov Illness-death Model

Luís Meira-Machado, Jacobo de Uña-Álvarez and Somnath Datta

Report 12/05

Discussion Papers in Statistics and Operation Research

Imprime: GAMESAL

Edita:

Universidade de Vigo
Facultade de CC. Económicas e Empresariales
Departamento de Estatística e Investigación Operativa
As Lagoas Marcosende, s/n 36310 Vigo
Tfno.: +34 986 812440

I.S.S.N: 1888-5756

Depósito Legal: VG 1402-2007

Conditional Transition Probabilities in a non-Markov Illness-death Model

Luís Meira-Machado[a], Jacobo de Uña-Álvarez[b], Somnath Datta[c]

[a] *Department of Mathematics and Applications*
University of Minho
Campus de Azurem, 4800-058 Guimarães, Portugal
Telephone: (+351) 253510443
Fax: (+351) 253510401
E-mail: lmachado@math.uminho.pt
[b] *Department of Statistics and O.R.*
University of Vigo, Spain.
[c] *Department of Bioinformatics and Biostatistics*
University of Louisville, Louisville, USA.

Abstract

One important goal in multi-state modeling is the estimation of transition probabilities. In longitudinal medical studies these quantities are particularly of interest since they allow for long-term predictions of the process. In recent years significant contributions have been made regarding this topic. However, most of the approaches assume independent censoring and do not account for the influence of covariates. This paper introduces feasible estimation methods for the transition probabilities in an illness-death model conditionally on current or past covariate measures. These approaches are evaluated through a simulation study, comparing two different estimators. The proposed methods are illustrated using real data.

Keywords: Conditional Survival, Dependent Censoring, Illness-death model, Kaplan-Meier, Multi-state model, Transition probabilities

1. Introduction

The so-called "illness-death" model plays a central role in the theory and practice of multi-state models (Andersen et al. [5], Meira-Machado et al. [16]). In the irreversible version of this model, individuals start in the "healthy" state and subsequently move either to the "diseased" state or to the "dead"

state. Individuals in the "diseased" state will eventually move to the "dead" state without any possibility of recovery. See Figure 1. Many time-to-event data sets from medical studies with multiple end points can be reduced to this generic structure. Thus, methods developed for the three-state illness-death model have a wide range of applications. From a theoretical standpoint, this is the simplest multi-state generalization of the survival analysis model that incorporates both branching (as in a multiple decrement/competing risk model) and an intermediate state (as in a progressive tracking model). Thus, unlike the survival or the competing risk model, this model is not necessarily Markovian.

Various aspects of the model dynamics are captured by the transition probabilities. In the presence of right censoring, these can be estimated by the Aalen-Johansen product limit estimator (Aalen and Johansen [1]) provided the system is Markovian. However, as demonstrated by Meira-Machado et al. [15], the Aalen-Johansen estimator is inconsistent when the Markov assumption does not hold. They also illustrate through a real data example that the Markovianity cannot be taken for granted. Meira-Machado et al. [15] and Amorim et al. [4] provide alternative nonparametric estimators specific to the three-state illness-death model that are consistent even without the Markov assumption.

In this paper, we revisit the problem of estimation of the transition probabilities of an irreversible, possibly non-Markov illness-death model. However, unlike the previous attempts, we are interested in a regression setup where we estimate these probabilities given a continuous covariate that could either be a baseline covariate or a current covariate that is observed for an individual before the individual makes a particular transition of interest. Our methodology is motivated by the colon cancer data set originally investigated by Moertel et al. [17] and subsequently reanalyzed by Lin et al. [14] to study the joint distribution of gap times between enrolment (curative surgery), the disease recurrence and death. These data can also be viewed as arising from a three-state illness-death model where "recurrence" can be modeled as the intermediate illness state. We are interested in the effect of a covariate (age at surgery, or number of lymph nodes with detectable cancer), on the probabilities of transitions between the several states. Standard regression models in this setup (besides of imposing Markovianity) usually rely on a parametric specification of the covariates' effects on the intensity functions; therefore, flexible effects of the covariates on the transition probabilities as those depicted in Figures 3 and 4 (Section 4) can not be estimated through standard

techniques.

Another illustrative example is provided by the Bone-Marrow transplant data (Copelan et al. [7]). In this data set, one major intermediate event of interest is the development of acute graft versus host disease (GVHD). The terminal state is "relapse of leukemia or death". We may combine all the earlier states into one initial state (cancer and cGVHD free living) and view this as a three-state illness-death model. An interesting question to ask in this context is does the time from bone marrow transplant to the onset of acute GVHD affect the transition probabilities to the terminal state? We return to these questions in Section 4.

We provide two competing nonparametric regression estimators of the transition probability matrix of a three-state progressive illness-death model. We show that both estimators are valid (e.g., consistent) under mid regularity conditions even when the system is non-Markov or conditionally non-Markov. In both estimators, local smoothing is done by introducing kernel weights that are either based on a local constant (i.e. Nadaraya-Watson) or a local linear regression. Right censoring is handled by appropriate reweighting of the chosen summands and the differences between the two estimators are somewhat subtle in this regard. The first estimator is based on observations that are completely uncensored (i.e., fully observed till death) whereas the second estimator is based on observations that were uncensored till a given time. Extensive simulation studies are provided comparing the two estimators.

The rest of the paper is organized as follows. Section 2 introduces the formal notations and the two estimators. Section 3 describes the simulation setup and the findings of a number of simulation experiments. Illustrative real data applications are provided in Section 4. The main body of the paper ends with a discussion section (Section 5). Additional simulation results are presented in the Appendix.

2. Conditional Transition Probabilities

2.1. Notation

A multi-state model is a stochastic process $(X(t), t \in \mathcal{T})$ with a finite state space, where $X(t)$ represents the state occupied by the process at time $t \geq 0$. In this paper we consider the progressive illness-death model depicted in Figure 1 and we assume that all the subjects are in state 1 at time

$t = 0$. This model is encountered in many medical studies (cancer studies, transplantations, etc) where State 1 is some initial stage of the disease (e.g. healthy, disease-free, etc), State 2 is some intermediate stage of the disease (e.g. alive with local recurrence, certain stage of a disease, transplantation, etc) and State 3 is an absorbing state (e.g. dead) which at some future time all subjects are expected to arrive. For this model the transitions allowed are $1 \to 2$, $1 \to 3$ and $2 \to 3$. This means that an individual may visit State 2 or going directly to State 3 without visiting State 2.

For two states i, j and two time points $s < t$, introduce the so-called transition probabilities

$$p_{ij}(s,t) = P\left(X(t) = j | X(s) = i\right).$$

In the illness-death model we only need to estimate three different transition probabilities: $p_{11}(s,t)$, $p_{12}(s,t)$, and $p_{22}(s,t)$. The two other transition probabilities ($p_{13}(s,t)$ and $p_{23}(s,t)$) can be obtained from these ones since $p_{13}(s,t) = 1 - p_{11}(s,t) - p_{12}(s,t)$ and $p_{23}(s,t) = 1 - p_{22}(s,t)$.

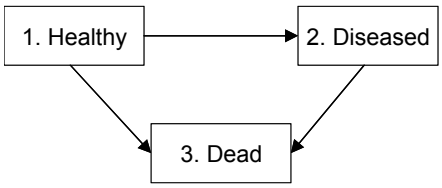

Figure 1: Illness-death model

In the framework of the progressive illness-death model, we may consider three random variables T_{12}, T_{13} and T_{23}, that represent the potential transition times from one state to another one. According to this notation, subjects not visiting state 2 will reach the absorbing state at time T_{13}. This time will be $T_{12} + T_{23}$ if he/she passes through state 2 before, where the variables T_{12} and T_{23} are recorded successively, rather than simultaneously. In this model we have two competing transitions leaving state 1. Therefore, we denote by $\rho = I(T_{12} \leq T_{13})$ the indicator of visiting state 2 at some time, $Z = T_{12} \wedge T_{13}$ the sojourn time in state 1, and $T = Z + \rho T_{23}$ the total survival time of the process.

Let C be the univariate censoring variable and put $\widetilde{Z} = Z \wedge C$ and $\widetilde{T} = T \wedge C$ for the censored versions of Z and T. Then, let $\Delta_1 = I(Z \leq C)$

and $\Delta = I(T \leq C)$ denote the respective censoring indicators. Note that ρ is observed only when $\Delta_1 = 1$. According to this notation, the transition probabilities may be written as

$$\begin{aligned} p_{11}(s,t) &= \frac{P\left(Z > t\right)}{P\left(Z > s\right)}, \;\; p_{12}(s,t) = \frac{P\left(s < Z \leq t, T > t\right)}{P(Z > s)} \\ p_{22}(s,t) &= \frac{P\left(Z \leq s, T > t\right)}{P\left(Z \leq s, T > s\right)}. \end{aligned}$$

Note that $p_{11}(s,t)$ and the denominator of $p_{12}(s,t)$ only involve the Z variable, while the remaining quantities involve expectations of particular transformations of the pair (Z,T), $S\left(\varphi\right) = E\left[\varphi\left(Z,T\right)\right]$. In Meira-Machado et al. [15] and Amorim et al. [4] the authors proposed to estimate these quantities using Kaplan-Meier weights pertaining to the distribution of the total time to weight the data. They showed that their estimator may behave much more efficiently than the competing ones, particularly when the underlying process is not Markovian. However, their methods are only valid if the censoring variable is assumed to be independent of the process. Furthermore, they do not account for the influence of covariates.

In this work we are interested in estimating the conditional transition probabilities: $p_{11}(s,t \mid X)$, $p_{12}(s,t \mid X)$, and $p_{22}(s,t \mid X)$ that can be computed for any times s and t, $s < t$ but conditional to some covariate value which we denote by X. Again, following the notation introduced above, the conditional transition probabilities are written as

$$\begin{aligned} p_{11}(s,t \mid X) &= \frac{P\left(Z > t \mid X\right)}{P\left(Z > s \mid X\right)}, \;\; p_{12}(s,t \mid X) = \frac{P\left(s < Z \leq t, T > t \mid X\right)}{P(Z > s \mid X)} \\ p_{22}(s,t \mid X) &= \frac{P\left(Z \leq s, T > t \mid X\right)}{P\left(Z \leq s, T > s \mid X\right)}. \end{aligned}$$

Now, the conditional transition probability $p_{11}(s,t \mid X)$ and the denominator of $p_{12}(s,t \mid X)$ only involve the conditional distribution of Z given X. This conditional distribution can be estimated nonparametrically following Beran [6]. The remaining quantities involve expectations of particular transformations of the pair (Z,T) given X, $S\left(\varphi \mid X\right) = E\left[\varphi\left(Z,T\right) \mid X\right]$ which can not be estimated so simply. In particular, we need to estimate the expectations $S\left(\varphi_{s,t} \mid X\right)$ and $S\left(\widetilde{\varphi}_{s,t} \mid X\right)$, where $\varphi_{s,t}(u,v) = I(s < u \leq t, v > t)$ and $\widetilde{\varphi}_{s,t}(u,v) = I(u \leq s, v > t)$.

In the following, we discuss how these expectations can be empirically approximated from the data $\left\{\left(\widetilde{Z}_i, \widetilde{T}_i, \Delta_{1i}, \Delta_i, \Delta_{1i}\rho_i, X_i\right), 1 \leq i \leq n\right\}$, which are assumed to form a random sample of the vector $\left(\widetilde{Z}, \widetilde{T}, \Delta_1, \Delta, \Delta_1\rho, X\right)$. We will estimate these quantities assuming that the censoring variable C is independent of (Z, T) given X. Note that this assumption does not exclude the possibility of dependent censoring (i.e., C unconditionally dependent on (Z, T)). Markovianity will not be assumed.

2.2. The Estimators

In this section, we will introduce two estimators for the conditional transition probabilities, $p_{hj}(s, t \mid X)$, in an illness-death model. Both methods are based on Inverse Probability of Censoring Weighted (Lin et al., 1999; Satten and Datta, 2001). As mentioned in Section 2.1, this can be done via estimating the general conditional expectation $E\left[\varphi\left(Z, T\right) \mid X = x\right]$. To estimate this quantity we may use kernel smoothing techniques by calculating a local average of the $\varphi(Z, T)$. This can be written as $\sum_{i=1}^{i=n} W_{1i}(x)\varphi(Z_i, T_i)$ where $W_{1i}(x)$ is a weight function which can be estimated using Nadaraya-Watson (Nadaraya [19],Watson [22]) or local linear estimators. In our case, we have to estimate $f(x) = E\left[\varphi_{s,t}\left(Z, T\right) \mid X = x\right]$, $g(x) = E\left[\widetilde{\varphi}_{s,t}\left(Z, T\right) \mid X = x\right]$ and $h(x) = E[\xi_s\left(Z\right) \mid X = x]$, where $\varphi_{s,t}\left(u, v\right) = I(s < u \leq t, v > t)$, $\widetilde{\varphi}_{s,t}(u, v) = I(u \leq s, v > t)$ and $\xi_s(u) = I(u > s)$.

To estimate these quantities, we need to estimate the d.f. of C given X, G_X. Let G_{X_i} denote the conditional distribution function of $C \mid X = X_i$ and let $\widehat{G}_{X_i}$ stand for its estimator. The estimation of the conditional distribution function of the response, given the covariate under random censoring has been considered in many papers. This topic was introduced by Beran [6] and was further studied by several authors (see e.g. papers by Dabrowska [8], Dabrowska [9], Dabrowska [10], Dabrowska [11]; Akritas [2]; Van Keilegom et al. [21] and Van Keilegom [20]). Recently, Beran's estimator has been extended to regression of state occupation probabilities of a multi-state model by Mostajabi and Datta [18].Their proposals can also be used to estimate the conditional distribution function of $C \mid X = x$, say $\widehat{G}_x$. This can be done using the estimator introduced by Beran [6],

$$\widehat{G}_x(t) = \prod_{T_i \leq t, \Delta_i = 0} \left[1 - \frac{W_{0i}(x, a_n)}{\sum_{j=1}^{n} I(T_j \geq T_i) W_{0j}(x, a_n)}\right] \tag{1}$$

with

$$W_{0i}(x, a_n) = \frac{K_0\left((x - X_i)/a_n\right)}{\sum_{j=1}^{n} K_0\left((x - X_j)/a_n\right)}$$

where $W_{0i}(x, a_n)$ are the Nadaraya-Watson (NW) weights, K_0 is a known probability density function (the kernel function) and a_n is a sequence of bandwidths. This estimator reduces to the so-known Kaplan-Meier (Kaplan and Meier [13]) estimator when all the weights are equal. To cope with left-truncated data, one can also use the estimator of the conditional distribution, proposed by Iglésias-Pérez and González-Manteiga [12].

In order to introduce our estimators note that, assuming that the support of the conditional distribution of T is contained in that of $C \mid X$, we have $E[\varphi(Z, T) \mid X] = E[\varphi(\widetilde{Z}, \widetilde{T})\Delta/(1 - G_X(\widetilde{T}^-)) \mid X)]$. We propose to plug-in Beran's estimator $\hat{G}_X$ and use NW or a local linear estimator (LLE) to estimate $f(x)$, i.e. to compute

$$\hat{f}(x; s, t) = \sum_{i=1}^{n} W_{1i}(x, b_n) \frac{\varphi_{s,t}(\widetilde{Z}_i, \widetilde{T}_i)\Delta_i}{1 - \hat{G}_{X_i}(\widetilde{T}_i^-)} = \sum_{i=1}^{n} W_{1i}(x, b_n) \frac{I(s < \widetilde{Z}_i \leq t, \widetilde{T}_i > t)\Delta_i}{1 - \hat{G}_{X_i}(\widetilde{T}_i^-)}$$

where $W_{1i}(x, b_n)$ are NW weights as introduced above, or using local linear weights,

$$W_{1i}(x, b_n) = \frac{K_1\left((x - X_i)/b_n\right)\left[S_{n,2}(x) - (x - X_i)S_{n,1}(x)\right]}{\sum_{j=1}^{n} K_1\left((x - X_j)/b_n\right)\left[S_{n,2} - (x - X_j)S_{n,1}(x)\right]}$$

with $S_{n,l} = \sum_{i=1}^{n} K_1((x - X_i)/b_n)(x - X_i)^l$, $l = 0, 1, 2$ and where b_n is a sequence of bandwidths and K_1 is a known kernel function.

Similarly, we can use Nadaraya-Watson estimators or local linear estimators to estimate $g(x)$ and $h(x)$ i.e.

$$\hat{g}(x; s, t) = \sum_{i=1}^{n} W_{1i}(x, b_n) \frac{\widetilde{\varphi}_{s,t}(\widetilde{Z}_i, \widetilde{T}_i)\Delta_i}{1 - \hat{G}_{X_i}(\widetilde{T}_i^-)} = \sum_{i=1}^{n} W_{1i}(x, b_n) \frac{I(\widetilde{Z}_i \leq s, \widetilde{T}_i > t)\Delta_i}{1 - \hat{G}_{X_i}(\widetilde{T}_i^-)}$$

and

$$\hat{h}(x;s) = \sum_{i=1}^{n} W_{1i}(x,c_n)\frac{\xi_s(\widetilde{Z}_i)\Delta_{1i}}{1-\hat{H}_{X_i}(\widetilde{Z}_i^-)} = \sum_{i=1}^{n} W_{1i}(x,c_n)\frac{I(\widetilde{Z}_i \geq s)\Delta_{1i}}{1-\hat{H}_{X_i}(\widetilde{Z}_i^-)}$$

where $\hat{H}_X$ stands for the Kaplan-Meier estimator of the conditional distribution of C given X based on the $(\widetilde{Z}_i, 1-\Delta_{1i})$'s.

Then, we may introduce Inverse Probability Censoring Weighted estimators (IPCW) for the conditional transition probabilities, as follows:

$$\hat{p}_{11}(x;s,t) = \hat{p}_{11}(s,t \mid X=x) = \frac{\hat{h}(x;t)}{\hat{h}(x;s)},$$
$$\hat{p}_{12}(x;s,t) = \hat{p}_{12}(s,t \mid X=x) = \frac{\hat{f}(x;s,t)}{\hat{h}(x;s)},$$
$$\hat{p}_{22}(x;s,t) = \hat{p}_{22}(s,t \mid X=x) = \frac{\hat{g}(x;s,t)}{\hat{g}(x;s,s)}.$$

Alternatively, by noting that $E[\varphi_{s,t}(Z,T) \mid X] = E[I(Z \leq s, T > t) \mid X] = E[I(Z \leq s, T > t)I(C > t)/(1-G_X(t^-)) \mid X]$, a different set of estimators may be introduced. This approach has been used previously by Lin et al. [14] to estimate the bivariate distribution for censored gap times. In our setup, alternative estimators of the transition probabilities will involve the following estimators:

$$\tilde{f}(x;s,t) = \sum_{i=1}^{n} W_{1i}(x,b_n)\frac{\varphi_{s,t}(\widetilde{Z}_i,\widetilde{T}_i)}{1-\hat{G}_{X_i}(t^-)} = \sum_{i=1}^{n} W_{1i}(x,b_n)\frac{I(s<\widetilde{Z}_i \leq t, \widetilde{T}_i > t)}{1-\hat{G}_{X_i}(t^-)}$$

$$\tilde{g}(x;s,t) = \sum_{i=1}^{n} W_{1i}(x,b_n)\frac{\widetilde{\varphi}_{s,t}(\widetilde{Z}_i,\widetilde{T}_i)}{1-\hat{G}_{X_i}(t^-)} = \sum_{i=1}^{n} W_{1i}(x,b_n)\frac{I(\widetilde{Z}_i \leq s, \widetilde{T}_i > t)}{1-\hat{G}_{X_i}(t^-)}$$

and

$$\tilde{h}(x;s) = \sum_{i=1}^{n} W_{1i}(x,c_n)\frac{\xi_s(\widetilde{Z}_i)\Delta_{1i}}{1-\hat{H}_{X_i}(s^-)} = \sum_{i=1}^{n} W_{1i}(x,c_n)\frac{I(\widetilde{Z}_i \geq s)\Delta_{1i}}{1-\hat{H}_{X_i}(s^-)}.$$

These lead to the conditional transition probabilities (LIN-based) given by

$$\tilde{p}_{11}(x;s,t) = \tilde{p}_{11}(s,t \mid X=x) = \frac{\tilde{h}(x;t)}{\tilde{h}(x;s)},$$
$$\tilde{p}_{12}(x;s,t) = \tilde{p}_{12}(s,t \mid X=x) = \frac{\tilde{f}(x;s,t)}{\tilde{h}(x;s)},$$
$$\tilde{p}_{22}(x;s,t) = \tilde{p}_{22}(s,t \mid X=x) = \frac{\tilde{g}(x;s,t)}{\tilde{g}(x;s,s)}.$$

Consistency and further asymptotics for the proposed estimators can be derived as usual. More focused in practical issues, the finite-sample performance of IPCW estimators and the alternative LIN-based estimators is investigated by simulations in the following section.

3. Simulation Study

In this section we carry out some simulations to investigate the behavior of the proposed estimators for finite sample sizes. More specifically, the estimators $\hat{p}_{11}(x;s,t)$, $\hat{p}_{12}(x;s,t)$, $\hat{p}_{22}(x;s,t)$, $\tilde{p}_{11}(x;s,t)$, $\tilde{p}_{12}(x;s,t)$ and $\tilde{p}_{22}(x;s,t)$ introduced in Section 2 are considered.

To simulate the data in the illness-death model, we follow closely the work described by Amorim et al. [4], but including covariate effects. In summary, the simulation procedure is as follows:

Step 1. Draw $\rho \sim Ber(p)$ where p is the proportion of subjects passing through State 2.

Step 2. If $\rho = 1$ then:

(2.1) $V_1 \sim U(0,1), V_2 \sim U(0,1)$ and $X \sim U(0,1)$ are independently generated;

(2.2) $U_1 = V_1, A = (2U_1 - 1) - 1, B = (1 - (2U_1 - 1))^2 + 4V_2(2U_1 - 1)$

(2.3) $U_2 = 2V_2 / \left(\sqrt{B} - A\right)$

(2.4) $Z = \ln(1/(1-U_1))$ and $\lambda(X) = 0.6X + 0.4$

(2.5) $Z(X) = Z/\lambda(X), T = \ln(1/(1-U_2)) + Z(X)$

If $\rho = 0$ then $Z = Z(X)$.

Situations with $p = 1$ corresponds to the three-state progressive model, in which a direct transition $1 \to 3$ is not allowed. In our simulation we

consider $p = 0.7$. To allow for dependent censoring, we consider a different scenario where $C|X = x$ is generated from a exponential distribution with rate $\lambda(x) = 0.15 + 0.35x$. This implies a censoring percentage of about 42%.

In Figure 2 we plot the IPCW and Lin-based conditional transition probabilities, by fixing $s = 0.2231$ and considering two possible values for the covariate information (first and third quartile). The results, which are estimators averaged along 1,000 Monte Carlo trials of size $n = 100$, show that (a) IPCW-type and Lin-based estimators are close to each other, and that (b) the transition probabilities greatly depend on covariate information, particularly $p_{11}(x; 0.2231, t)$ and $p_{12}(x; 0.2231, t)$ (not so clear for $p_{22}(x; 0.2231, t)$). This influence of the covariate can be also seen from the simulation steps described above: larger values of X are associated to smaller sojourn times in state 1 and, consequently, to a smaller survival (T).

The aim of this simulation study is to investigate the performance of the two proposed estimators (IPCW and LIN-based) and to compare them to each other. For measuring the estimates' performance, we computed the integrated mean square error (IMSE) of the estimates. For each simulated setting we derived the analytic expression of $p_{ij}(x; s, t)$ so the MSE of the estimator could be computed. K = 1000 Monte Carlo trials were generated, with two different sample sizes $n = 100$ and $n = 200$. Let $\hat{p}_{ij}^k(x; s, t)$ denote the estimated conditional transition probability based on the kth generated data set. For each fixed (x, s, t) we computed the pointwise estimates of the MSE as:

$$\widehat{MSE}(\hat{p}_{ij}(x; s, t)) = \frac{1}{K} \sum_{k=1}^{K} [\hat{p}_{ij}^k(x; s, t) - p_{ij}(x; s, t)]^2 \tag{2}$$

To summarize the results we fixed the values of (s, t) using several quantiles (the same pairs as those used in the paper by Lin et al. [14]) and we calculated the IMSE as

$$\widehat{IMSE} = \sum_{x_l} \widehat{MSE}(\hat{p}_{ij}(x_l; s, t)) \times \delta \tag{3}$$

where x_l denotes a set of grid points for the covariate, going from 0 to 1 with step $\delta = 0.025$. The results are displayed in Tables 1 to 3. To compute the conditional transition probabilities $\hat{p}_{ij}(x; s, t)$ and $\tilde{p}_{ij}(x; s, t)$ we have used a common bandwidth selector and Gaussian kernels. To this end we have used the `dpik` function which is available from the `R` `KernSmooth` package. This

is the data based bandwidth selector of Wand and Jones (1995). For the computation of $W_1(x; b_n)$ we have used Nadaraya-Watson (NW) and local linear weights (for the weights W_0 of the Beran's estimator we simply used NW). Since the results for NW weights were always superior to those based on local linear weights, we only provide here the results corresponding to the former. Additional simulation results are provided in the Appendix.

When using NW weights the two estimators (IPCW and LIN-based) for $p_{11}(x; s, t)$ are equal and, therefore, in Table 1 we only give one set of results. In general, both methods provide good results with IMSE values which decrease with an increasing sample size. It is also seen that the estimation of the transition probabilities is performed with less accuracy as s and t grow but for $p_{22}(x; s, t)$, for which the smallest values of IMSE are obtained for large s and t. Results shown in Table 2 suggest that the IPCW method leads to better results for $p_{12}(x; s, t)$ while the contrary occurs when estimating $p_{22}(x; s, t)$ (Table 3). Therefore, no one of the proposed estimators seems to be uniformly the best.

		t	0.5108	0.9163	1.6094
	s				
n=100	0.2231		5.2206	8.8258	10.0934
	0.5108		—	8.5920	13.3747
	0.9163		—	—	18.8724
n=200	0.2231		3.5581	5.9610	6.9856
	0.5108		—	6.1066	9.3146
	0.9163		—	—	12.6278

Table 1: IMSE ($\times$10000) of the estimated transition probabilities $\hat{p}_{11}(x; s, t)$ along 1,000 trials for different sample sizes

			t	0.5108	0.9163	1.6094
		s				
n=100	IPCW	0.2231		3.2962	5.3945	7.3948
	LIN-based			3.4047	5.7230	7.7928
	IPCW	0.5108		—	5.6419	9.8266
	LIN-based			—	5.9080	10.3677
	IPCW	0.9163		—	—	12.8726
	LIN-based			—	—	13.3364
n=200	IPCW	0.2231		2.1506	3.5847	4.9058
	LIN-based			2.2024	3.7851	5.1341
	IPCW	0.5108		—	3.8685	6.4965
	LIN-based			—	4.0774	6.9215
	IPCW	0.9163		—	—	8.6680
	LIN-based			—	—	9.0572

Table 2: IMSE ($\times$10000) of the estimated transition probabilities $\hat{p}_{12}(x; s, t)$ along 1,000 trials for different sample sizes

			t	0.5108	0.9163	1.6094
		s				
n=100	IPCW	0.2231		95.9706	92.1324	67.9388
	LIN-based			90.1238	80.4370	34.2330
	IPCW	0.5108		—	71.0090	60.3021
	LIN-based			—	65.9688	49.8399
	IPCW	0.9163		—	—	70.8846
	LIN-based			—	—	63.2712
n=200	IPCW	0.2231		71.8374	68.8493	48.0180
	LIN-based			63.7254	60.1681	29.1906
	IPCW	0.5108		—	51.8131	41.8810
	LIN-based			—	45.2639	33.6916
	IPCW	0.9163		—	—	52.2701
	LIN-based			—	—	45.1294

Table 3: IMSE ($\times$10000) of the estimated transition probabilities $\hat{p}_{22}(x; s, t)$ along 1,000 trials for different sample sizes

4. Example of Application

To illustrate our estimators we consider two real data sets. One of these data sets comes from the well-known colon cancer study which is freely available as part of the `R survival` package (Moertel et al. [17]). In addition to this data set we also use data on 137 Bone-Marrow transplant patients for leukemia (Copelan et al. [7]). In both data sets, non-fatal events (recurrence and the development of acute graft-versus-host disease, respectively) are observed during the disease course. These intermediate events can be modeled using the progressive illness-death model depicted in Figure 1.

4.1. Colon cancer data

These are data from one of the first successful trials of adjuvant chemotherapy for colon cancer. From the total of 929 patients, affected by colon cancer, that underwent a curative surgery for colorectal cancer, 468 developed a recurrence and among these 414 died. 38 patients died without recurrence. The remaining 423 patients contributed with censored survival times. For each individual, an indicator of his/her final vital status (censored or not), the survival times (time to recurrence, time to death) from the entry of the patient in the study (in days), and a vector of covariates including *age* (in years), *nodes* (number of lymph nodes with detectable cancer) and *recurrence* (coded as 1 = yes; 0 = no) were recorded. The covariate *recurrence* is a time-dependent covariate which can be used to identify an intermediate event in an illness-death model with states "Alive and disease-free", "Alive with recurrence" and "dead".

Using a Cox proportional hazards model, we verified that the transition rate from state 2 to state 3 is affected by the time spent in the previous state (p-value < 0.001). This allowed us to conclude that the Markov assumption may be unsatisfactory for the colon cancer data set and that, consequently, Aalen-Johansen type estimators should not be used. In this section we will present estimated transition probabilities conditionally on current or past covariate measures such as *age* or *nodes* (minimum = 0 and maximum = 13). These estimators were calculated using the IPCW method and/or LIN-based procedures as explained above. Both approaches do not assume the process to be Markovian, allowing for dependent censoring and flexible (i.e. nonparametric) covariate effects otherwise.

Figures 3 and 4 depict respectively the IPCW estimates of $p_{11}(x; 379, 1000)$ and $p_{12}(x; 379, 1000)$ as functions of the covariate *age* together with a 95%

pointwise confidence bands based on simple bootstrap which resamples each datum with probability $1/n$. In both plots it is seen that these curves are not constant; the effects of *age* depicted in these plots, which are purely nonparametric, indicate the real influence of this covariate in the survival prognosis. In fact, it would not be possible to include an horizontal line within the confidence bands of Figure 4, suggesting a significative influence of age on survival. More specifically, patients near forties have a larger probability of recurrence than older patients. This is in agreement with Figure 5 where it is shown, among other things, that 40 years old patients have a higher probability of recurrence than patients with 68 years (bottom-left plot). In Figure 6 we present similar plots for the covariate *nodes*, revealing that this covariate has also a real impact on the conditional transition probabilities.

Figures 7 and 8 report the results corresponding to the Lin-based estimator. Roughly speaking, conclusions from these plots are similar to those obtained from Figures 5 and 6. However, a particular problem of Lin-based estimator is appreciated at the bottom-left plots of Figures 7 and 8, because the displayed curves for $p_{22}(x; s, t)$ are not monotone decreasing in t and, therefore, they are not admissible. This is a consequence of the specific reweighting of the data which is used in this approach, which may lead to problems of interpretation at the right tail of the distribution.

4.2. Bone-Marrow transplant data

Bone-Marrow transplant data are discussed in (Copelan et al. [7]). In this data set, one intermediate event of major interest is the development of acute graft-versus-host disease (GVHD). Therefore, this data can also be viewed as an illness-death model (Figure 1) where "relapse of leukemia or death" is the absorbing state, GVHD is modeled as the intermediate illness state, and earlier states are combined into one initial state (cancer and GVHD free living).

The development of acute GVHD typically occurs within the first three months following transplantation and may change the patient's final prognosis. Therefore, it is important to know how the time from bone marrow transplant to the onset of acute GVHD affect the transition probability to the absorbing state. In Figure 9 we show the plots for the two proposed estimators of the transition probability $p_{22}(x; 120, t)$ for two times of GVHD ($x = 84$ and $x = 119$). Results show few differences between the two curves, particularly when looking at the IPCW estimator (left panel). This application is also interesting because it illustrates that the proposed methods can handle covariates which are not defined baseline (time to cGVHD is defined only for those going through state 2). The absence of differences between the depicted curves could be also interpreted as lack of evidence against Markovianity for these data set. In Figure 10 we show similar plots for $p_{22}(x; 122, t)$ according to age (again using first and third quantiles). In this case the two curves obtained for different ages separate, indicating a poorer survival prognosis for elderly people. Note that, as for the colon cancer data example, Lin-based estimator provides a non-monotone curve which is not admissible; the IPCW method may be preferable in practice due to this issue.

5. Conclusions and final remarks

There have been several recent contributions for the estimation of the transition probabilities in the context of multi-state models. However, most of the approaches assume independent censoring and do not account for the influence of covariates. In this paper we have proposed two estimation methods for the transition probabilities given a continuous covariate. Both methods are based on local smoothing which is introduced using regression weights. Two different schemes of inverse censoring probability reweighting have been used to deal with right censoring. In one approach, the corresponding estimator (reweighting) is based on observations that are fully observed till death, whereas the other estimator is based on observations that were uncensored till a given time.

We have investigated the performance of the estimators through simulations, showing that they are valid even when the system is non-Markov or conditionally non-Markov. None of the two proposed methods seem to dominate the other in all the possible scenarios. We have illustrated the proposed methodology using two real data sets. In particular, it has been illustrated that the introduced estimators may handle covariates which are not defined baseline. Besides, one of the two approaches (Lin-based one) has the drawback of occasionally providing non-monotone curves for transition probabilities which are indeed monotone and, therefore, its practical use could be less recommended.

An interesting open question is if this idea can be generalized (and how) to more complex multi-state models; this is left to future research. Another issue is the application of the proposed methods to multiple covariates. Although this could be formally done, the practical performance of the estimators heavily depend on the dimensionality. The presence of a moderate or large set of factors could recommend the application of some semiparametric technique to avoid the curse of dimensionality. Feasible solutions to this problem will be explored in the future.

Acknowledgements. This research was financed by FEDER Funds through Programa Operacional Factores de Competitividade COMPETE and by Portuguese Funds through FCT - Fundação para a Ciência e a Tecnologia, within Projects Est-C/MAT/UI0013/2011 and PTDC/MAT/104879/2008. We also acknowledge financial support from the project Grants MTM2008-03129 and MTM2011-23204 (FEDER support included) of the Spanish Ministerio de

Ciencia e Innovación and 10PXIB300068PR of the Xunta de Galicia. Partial support from a grant from the US National Security Agency (H98230-11-1-0168) is greatly appreciated.

References

[1] Aalen, O. and Johansen, S. [1978]. An empirical transition matrix for non homogeneous markov and chains based on censored observations, *Scandinavian Journal of Statistics* **5**: 141–150.

[2] Akritas, M. [1994]. Nearest neighbor estimation of a bivariate distribution under random censoring, *Annals of Statistics* **22**: 1299–1327.

[3] Altman, N. and Leger, C. [1995]. Bandwidth selection for kernel distribution function estimation, *Journal of Statistical Planning and Inference* **46**: 195–214.

[4] Amorim, A., de Uña-Álvarez, J. and Meira-Machado, L. [2011]. Presmoothing the transition probabilities in the illness-death model, *Statistics & Probability Letters* **81(7)**: 797–806.

[5] Andersen, P., Borgan, Ø., Gill, R. and Keiding, N. [1993]. *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.

[6] Beran, R. [1981]. *Nonparametric regression with randomly censored survival data*, Technical report, University of California, Berkeley.

[7] Copelan, E., Biggs, J. and Thompson, J. e. a. [1991]. Treatment for acute myelocytlic leukemia with allogeneic bone marrow transplantation following preperation with bu/cy, *Blood* **78**: 838–843.

[8] Dabrowska, D. [1987]. Non-parametric regression with censored survival data, *Scandinavian Journal of Statistics* **14**: 181–197.

[9] Dabrowska, D. [1988a]. Kaplan-meier estimate on the plane, *Annals of Statistics* **16**: 1475–1488.

[10] Dabrowska, D. [1988b]. Kaplan-meier estimate on the plane: weak convergence, lil, and the bootstrap, *Journal of Multivariate Analysis* **29**: 308–325.

[11] Dabrowska, D. [1989]. Uniform consistency of the kernel conditional kaplan-meier estimate, *Annals of Statistics* **17**: 1157–1167.

[12] Iglésias-Pérez, C. and González-Manteiga, W. [2003]. Bootstrap for the conditional distribution function with truncated and censored data, *The Annals of the Institute of Statistical Mathematics* **55**: 331–357.

[13] Kaplan, E. and Meier, P. [1958]. Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**: 457–481.

[14] Lin, D., Sun, W. and Ying, Z. [1999]. Nonparametric estimation of the gap time distributions for serial events with censored data, *Biometrika* **86**: 59–70.

[15] Meira-Machado, L., de Uña-Álvarez, J. and Cadarso-Suárez, C. [2006]. Nonparametric estimation of transition probabilities in a non-markov illness-death model, *Lifetime Data Analysis* **12**: 325–344.

[16] Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C. and Andersen, P. [2009]. Multi-state models for the analysis of time to event data, *Statistical Methods in Medical Research* **18**: 195–222.

[17] Moertel, C., Fleming, T. and McDonald, J.S., e. a. [1990]. Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma, *New England Journal of Medicine* **322**: 352–358.

[18] Mostajabi, F. and Datta, S. [2012]. Nonparametric regression of state occupation, entry, exit and waiting times with multistate right censored data, *Preprint* .

[19] Nadaraya, E. [1965]. On nonparametric estimates of density functions and regression curves, *Theory of Applied Probability* **10**: 186–190.

[20] Van Keilegom, I. [2004]. A note on the nonparametric estimation of the bivariate distribution under dependent censoring, *Journal of Nonparametric Statistics* **16**: 659–670.

[21] Van Keilegom, I., Akritas, M. and Veraverbeke, N. [2001]. Estimation of the conditional distribution in regression with censored data: a comparative study, *Computational Statistics and Data Analysis* **35**: 487–500.

[22] Watson, G. [1964]. Smooth regression analysis, *Sankhya* **26:15**: 175–184.

6. Appendix

In this Section we give the additional simulation results for the two estimators (IPCW and LIN-based) using local linear weights instead of NW weights. The results were obtained using the `dpik` function which is available from the `R KernSmooth` package. See Tables 4 to 6 below. We also performed additional simulations using other bandwidth selectors; for example, the plug-in bandwidth of Altman and Leger [3], `ALbw`, available from the `R kerdiest` package, was also used. This alternative bandwidth did not provide better results (not shown). Results for independent censoring were also obtained (not shown), leading to similar conclusions to those shown in Section 4 and in this Appendix.

			t	0.5108	0.9163	1.6094
		s				
n=100	IPCW	0.2231		5.5797	9.4730	10.9545
	LIN-based			5.6912	9.5955	11.0527
	IPCW	0.5108		—	9.3272	14.6103
	LIN-based			—	9.9418	14.7053
	IPCW	0.9163		—	—	20.7806
	LIN-based			—	—	20.9202
n=200	IPCW	0.2231		4.0298	6.7622	8.1146
	LIN-based			4.1739	6.9137	8.2597
	IPCW	0.5108		—	6.9762	10.8897
	LIN-based			—	7.1208	11.0621
	IPCW	0.9163		—	—	15.0724
	LIN-based			—	—	15.2672

Table 4: IMSE ($\times$10000) of the estimated transition probabilities $\hat{p}_{11}(x; s, t)$ along 1,000 trials for different sample sizes. Estimates based on the local linear estimators.

			t	0.5108	0.9163	1.6094
		s				
n=100	IPCW	0.2231		3.5132	5.7993	7.8780
	LIN-based			3.6389	6.1431	8.4336
	IPCW	0.5108		—	6.0819	10.3063
	LIN-based			—	6.4034	11.1962
	IPCW	0.9163		—	—	13.5681
	LIN-based			—	—	14.5392
n=200	IPCW	0.2231		2.4565	4.0349	5.6119
	LIN-based			2.5194	4.2707	5.9026
	IPCW	0.5108		—	4.3742	7.3794
	LIN-based			—	4.6275	7.9227
	IPCW	0.9163		—	—	9.9804
	LIN-based			—	—	10.4586

Table 5: IMSE ($\times$10000) of the estimated transition probabilities $\hat{p}_{12}(x; s, t)$ along 1,000 trials for different sample sizes. Estimates based on the local linear estimators.

			t	0.5108	0.9163	1.6094
		s				
n=100	IPCW	0.2231		99.1081	96.2793	71.6886
	LIN-based			95.5094	84.3036	35.2909
	IPCW	0.5108		—	74.8428	64.7501
	LIN-based			—	71.3421	52.7878
	IPCW	0.9163		—	—	74.8350
	LIN-based			—	—	68.3035
n=200	IPCW	0.2231		77.2416	74.7365	53.6386
	LIN-based			71.5834	66.4441	31.3221
	IPCW	0.5108		—	57.4873	48.5003
	LIN-based			—	52.4128	38.2466
	IPCW	0.9163		—	—	58.4093
	LIN-based			—	—	52.1681

Table 6: IMSE ($\times$10000) of the estimated transition probabilities $\hat{p}_{22}(x; s, t)$ along 1,000 trials for different sample sizes. Estimates based on the local linear estimators.

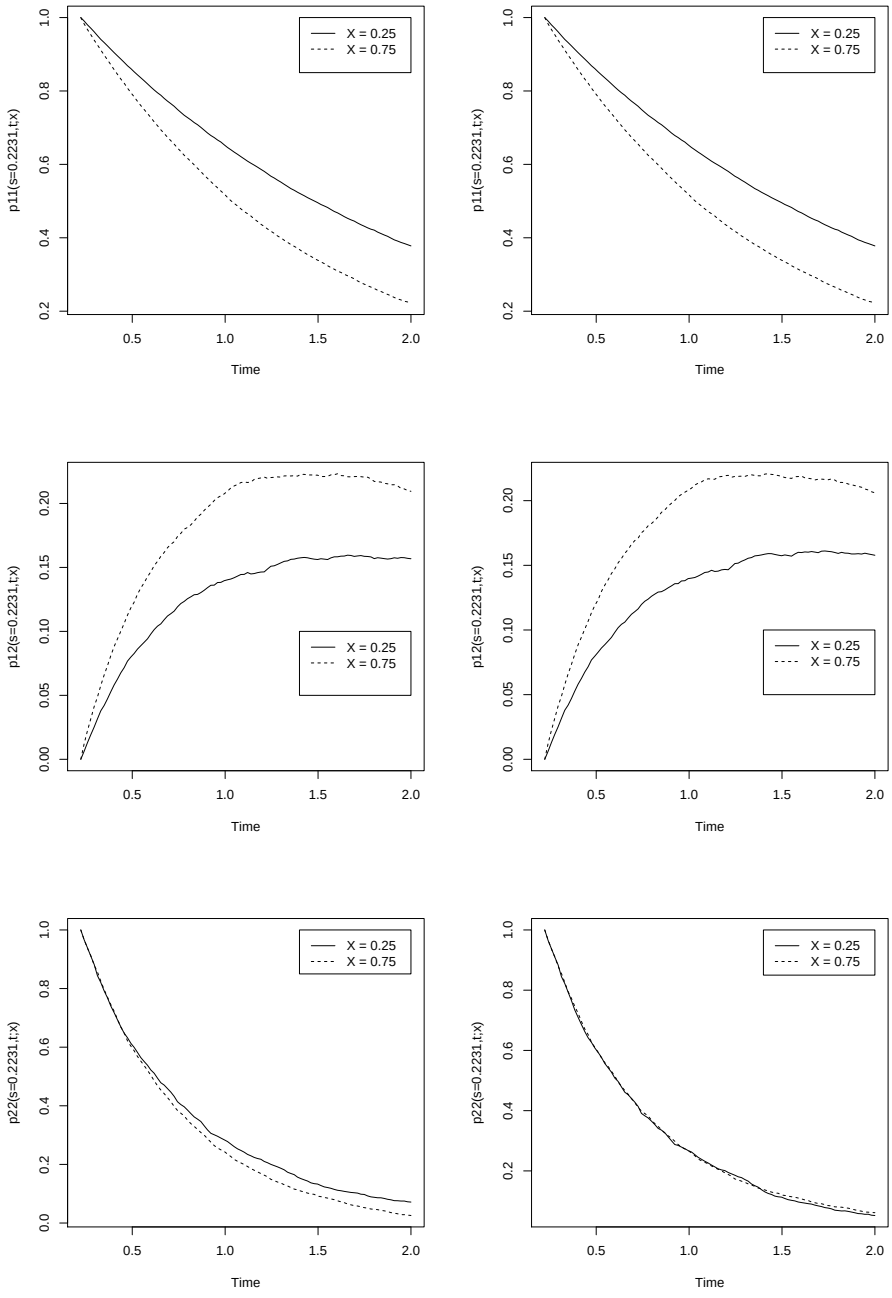

Figure 2: Conditional transition probabilities $P_{hj}(s,t;X)$ based on simulated data. IPCW method (left hand-side) and Lin's methods (right hand-side)

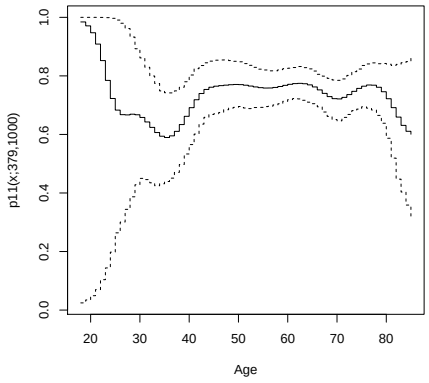

Figure 3: Evolution of the transition probability $p_{11}(379, 1000)$ along the covariate age with 95% bootstrap confidence bands. Colon cancer data.

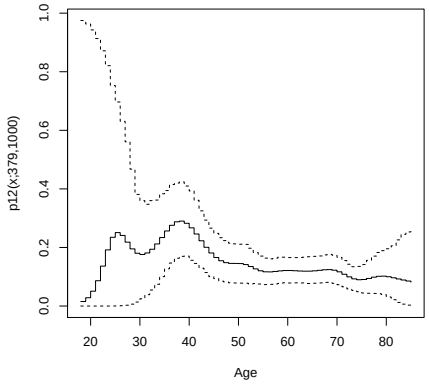

Figure 4: Evolution of the transition probability $p_{12}(379, 1000)$ along the covariate age with 95% bootstrap confidence bands (IPCW method). Colon cancer data.

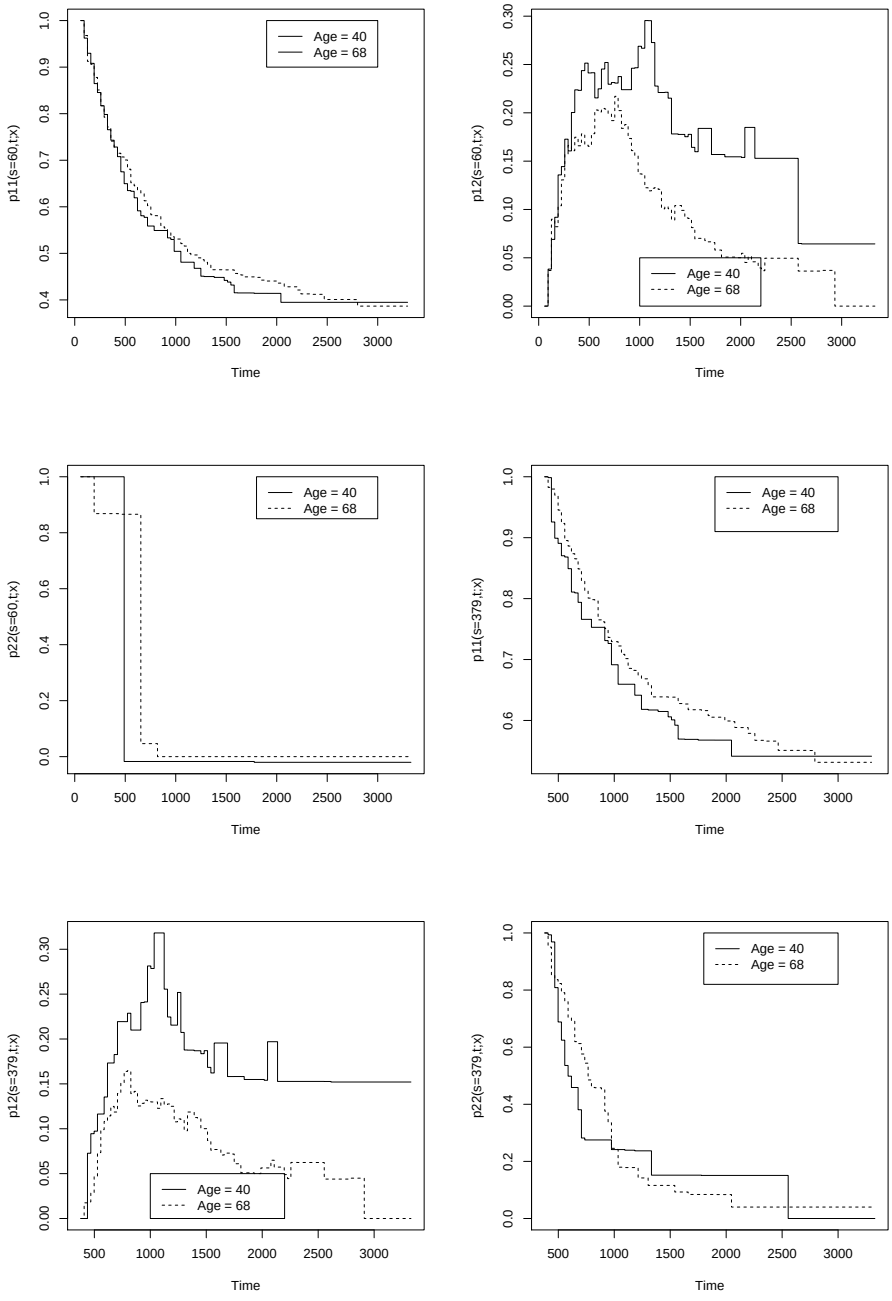

Figure 5: Conditional transition probabilities for the colon cancer data (IPCW method) for $age = 40$ and $age = 68$.

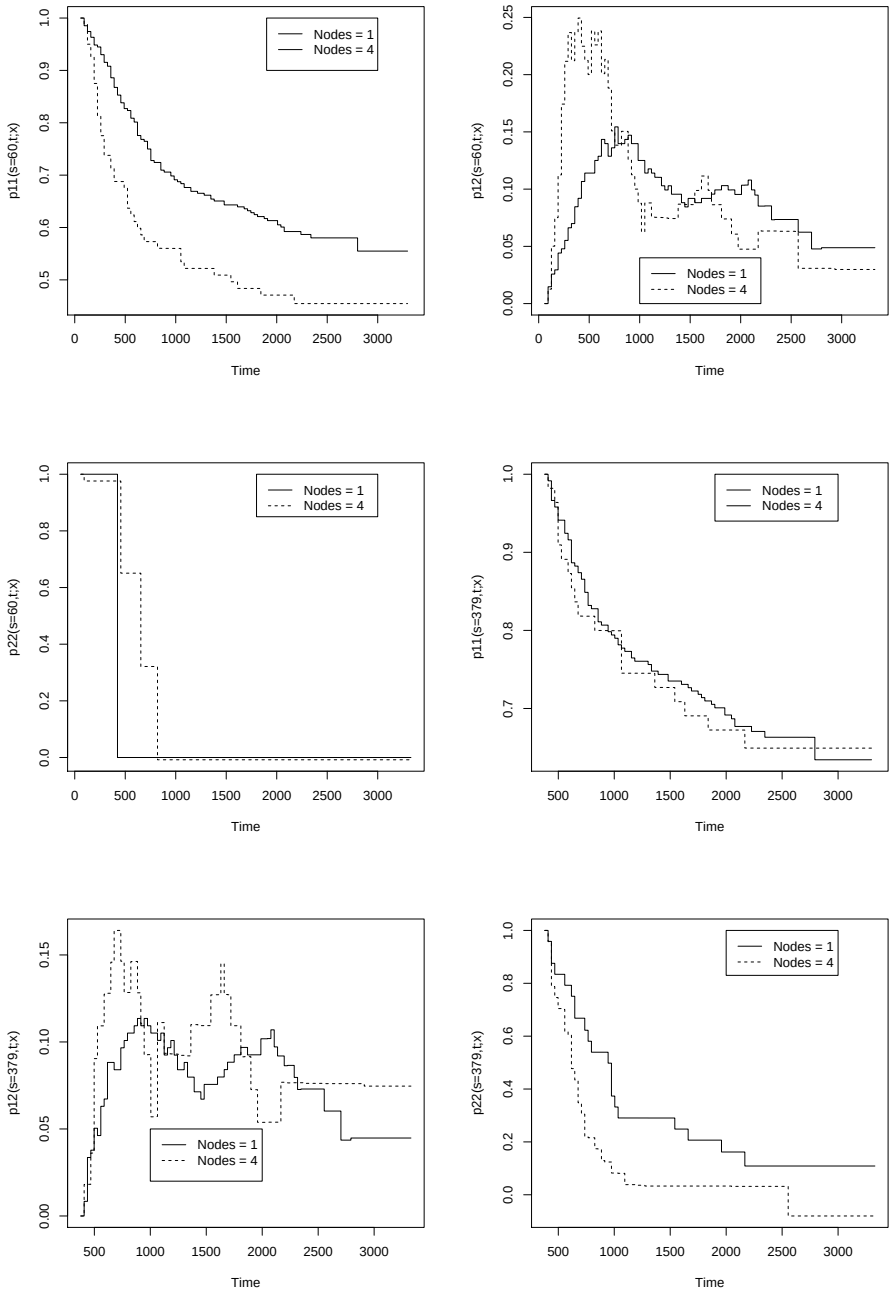

Figure 6: Conditional transition probabilities for the colon cancer data (IPCW method) for $nodes = 1$ and $nodes = 4$.

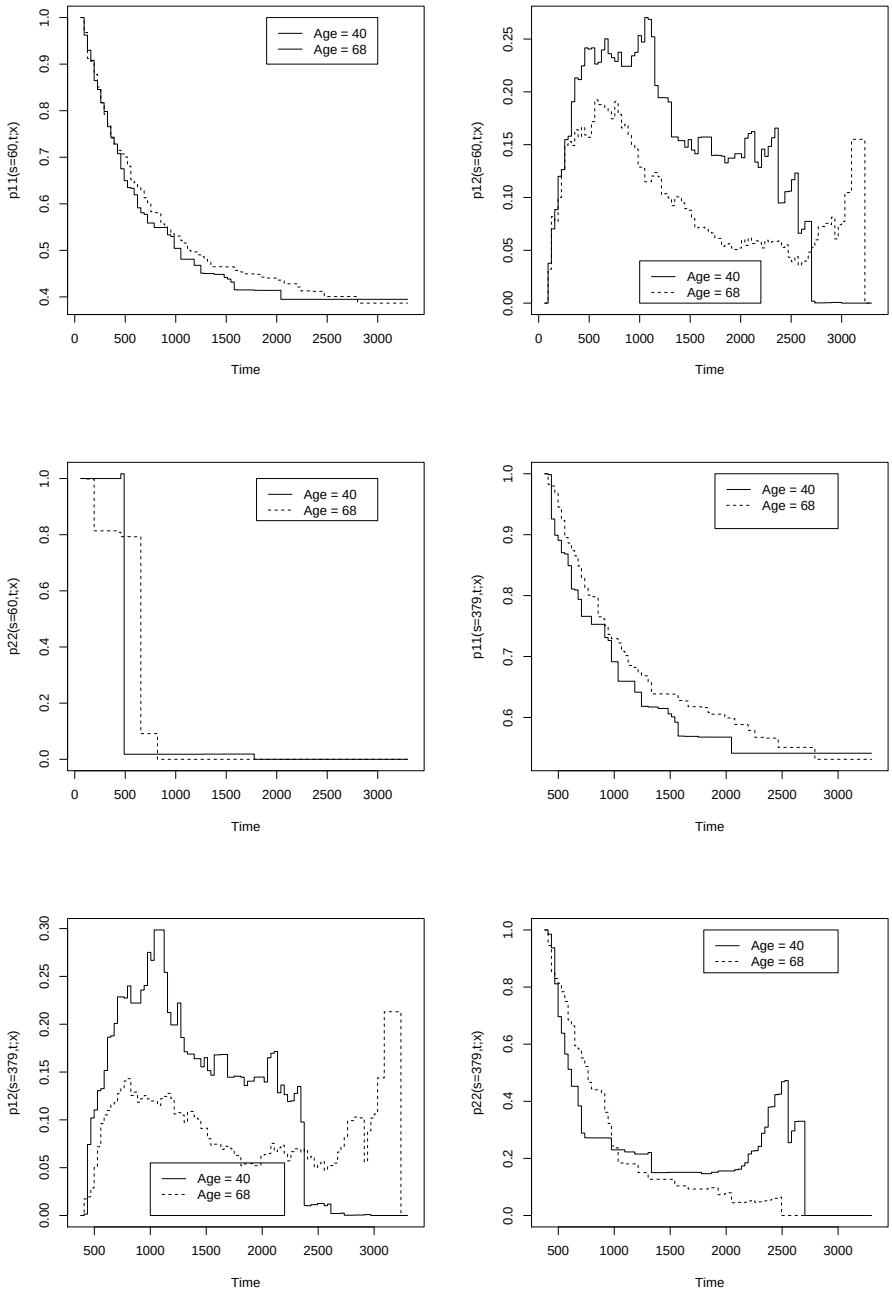

Figure 7: Conditional transition probabilities for the colon cancer data (LIN-based method) for $age = 40$ and $age = 68$.

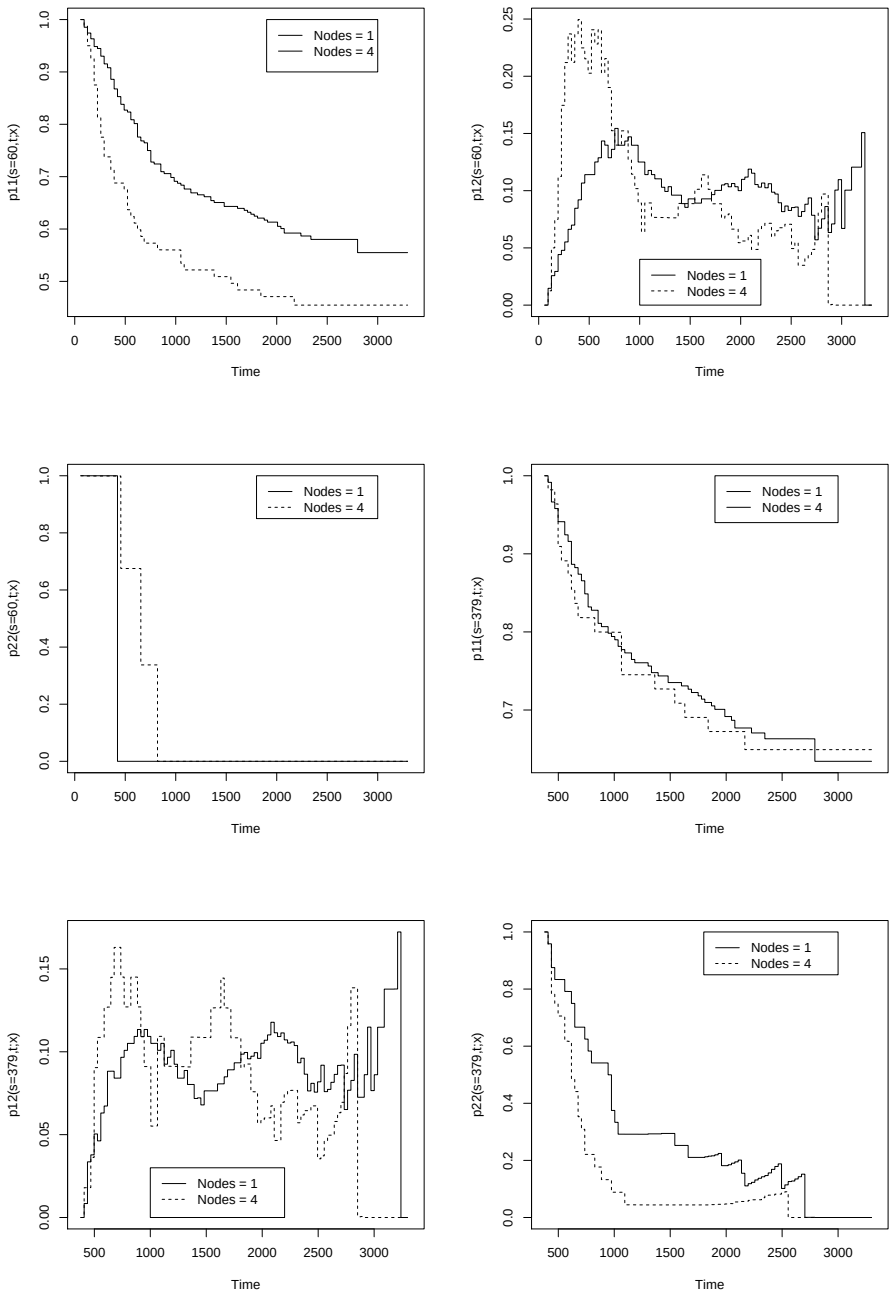

Figure 8: Conditional transition probabilities for the colon cancer data (LIN-based method) for $nodes = 1$ and $nodes = 4$.

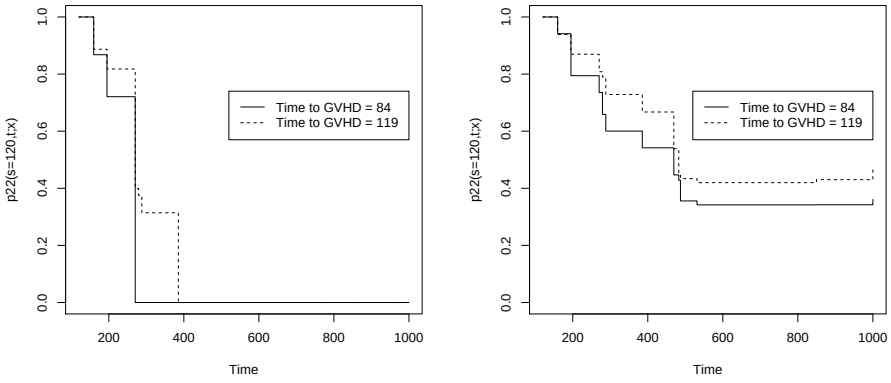

Figure 9: Conditional transition probabilities for the Bone-Marrow transplant data according to time to GVHD (left hand side - IPCW; right hand side - LIN-based method)

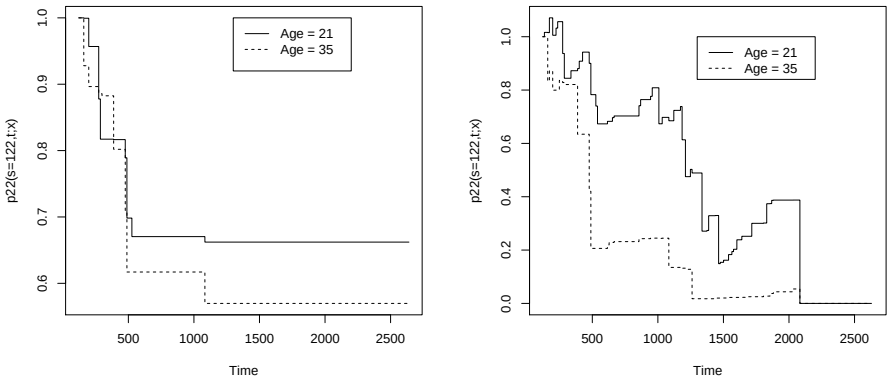

Figure 10: Conditional transition probabilities for the Bone-Marrow transplant data according age (left hand side - IPCW; right hand side - LIN-based method)