Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

Estimation of Transition Probabilities for the Illness-Death Model: Package TP.idm

Vanesa Balboa
Universidade de Vigo

Jacobo de Uña-Álvarez
Universidade de Vigo

Abstract

In this paper the R package **TP.idm** to compute an empirical transition probability matrix for the illness-death model is introduced. This package implements a novel non-parametric estimator which is particularly well suited for non-Markov processes observed under right censoring. Variance estimates and confidence limits are also implemented in the package.

Keywords: Aalen-Johansen, markov condition, multi-state models, nonparametric estimation.

1. Introduction

Multi-state models (Andersen, Borgan, Gill, and Keiding 1993; Commenges 1999; Hougaard 1999, 2000; Meira-Machado, de Uña-Álvarez, Cadarso-Suárez, and Andersen 2009) are the most commonly used models to describe longitudinal failure time data. A multi-state model is a model for a stochastic process $\{X(t), t \geq 0\}$ with a finite state space $S = \{1, \ldots, N\}$. In biomedical applications, the states may describe conditions like healthy, diseased, or clinical symptoms, or they might be based on biological markers or some scale of a given disease. A change of state is called a transition, or an event. States out of which transitions are modeled are called "transient states"; in contrast, "absorbing states" are states out of which transitions are not possible. The multi-state model is called progressive when the maximum number of visits to each state is one.

The multi-state process is fully characterized through transition probabilities between states h and j, defined for $0 \leq s < t$ as

$$\mathsf{P}_{hj}(s,t) = \mathsf{P}\left(X(t) = j | X(s) = h, H_{s^-}\right) \tag{1}$$

or through transition intensities

$$\alpha_{hj}(t) = \lim_{\Delta t \to 0} \frac{\mathsf{P}_{hj}\left(t, t + \Delta t\right)}{\Delta t} \tag{2}$$

representing the instantaneous hazard of progression to state j conditionally on occupying state h. The transition probability $\mathsf{P}_{hj}(s,t)$ depends in general on the evolution of the process over time, a "history" H_{s^-}, which collects the information on the process along the interval $[0, s)$. When the influence of H_{s^-} in (1) vanishes, the process is said to fulfill the Markov condition. Thus, Markov processes have history-free transition intensities (2), and they are characterized by a memoryless property, so the future evolution of the process just depends on its current state and the time elapsed since time origin, being independent of the states previously visited and the transition times among them. The Markov condition can be exploited both to investigate theoretical properties of the multi-state process and to derive estimation procedures (Andersen *et al.* 1993).

The progressive illness-death model is a particular multi-state model with three states, see Section 2 for details. The model is useful to describe the progress of individuals from an initial state to a terminal (absorbing) state, when they may undergo a certain intermediate event. Despite its relative simplicity, the illness-death or disability model (Hougaard 2000) has been widely used in the medical literature to describe the course of a disease, and to study the possible influence of the intermediate event on the probability of death. In Section 4 we apply the illness-death model to study progression after surgery for colon cancer patients, who may experience a recurrence or death during the follow-up. Special submodels of the illness-death model are the three-states progressive model (often used to analyze recurrent events), in which all the individuals undergo the intermediate state, or the competing risks model, in which the intermediate state becomes absorbing.

Several R packages for multi-state survival analysis are available on the Comprehensive R Archive Network (CRAN). For example, the **msSurv** package (Ferguson, Datta, and Brock 2012) provides nonparametric estimation for right censored, left truncated time to event data. This package can be used to estimate the state occupation probabilities $\mathsf{P}_{hj}(0,t)$ along with the corresponding variance estimates and confidence limits. The package **mvna** (Allignol, Beyersmann, and Schumacher 2008) computes the Nelson-Aalen estimator of the cumulative transition hazard, possibly subject to right censoring and left truncation. The package **etm** (Allignol, Schumacher, and Beyersmann 2011) provides the Aalen-Johansen transition probability matrix for a general multi-state model. It also features a Greenwood-type estimator of the covariance matrix, described in Andersen *et al.* (1993), Equation 4.4.17. The package handles both left truncation and right censored data. The **mstate** package (de Wreede, Fiocco, and Putter 2011) permits the estimation of transition probabilities with the Aalen-Johansen technique, possibly depending on covariates. The package can be applied to right censored and left truncated data in semiparametric or nonparametric multi-state models. The **msm** package (Jackson 2011) can be used to obtain estimates of the transition probabilities in continuous-time Markov and hidden Markov multi-state models for longitudinal data. Both options can be modeled in terms of covariates. The **p3state.msm** package (Meira-Machado and Pardiñas 2011) provides nonparametric estimates in the right censored progressive illness-death model, implementing the methods in Meira-Machado, de Uña-Álvarez, and Cadarso-Suárez (2006) as well as Cox-like regression models for the transition intensities. Later, Araújo, Meira-Machado, and Roca-Pardiñas (2014) developed the package **TPmsm**, which permits to

compute the Aalen-Johansen estimator in the illness-death model, and alternative transition probabilities estimates. These alternative estimators include presmoothed semiparametric estimators, and estimators which incorporate covariate effects by means of kernel smoothing too.

This paper describes the R package **TP.idm** (from *transition probabilities for the illness-death model*; Balboa-Barreiro, de Una-Alvarez, and Meira-Machado 2016) which is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=TP.idm`. This package implements a novel non-Markovian estimator for the transition probability matrix in the progressive illness-death model under right censoring, which has been proved to perform better than previously proposed methods (de Uña-Álvarez and Meira-Machado 2015). The package **TP.idm** reports empirical standard errors and confidence limits too. The new estimator follows the seminal idea of Pepe (1991), see also Pepe, Longton, and Thornquist (1991), being close to an estimator independently introduced by Titman (2015) too. For completeness, the Aalen-Johansen estimator and its standard error and confidence limits are also implemented.

The rest of this paper is organized as follows: Section 2 introduces the two aforementioned estimators for the transition probabilities, as well as different methods to compute confidence intervals. Section 3 briefly describes the **TP.idm** package. Section 4 illustrates the package through a real data application and, finally, Section 5 gives some concluding remarks.

2. Methods

The progressive illness-death model involves three states, $\{1, 2, 3\}$ say, and three possible transitions among them: $1 \to 2$, $1 \to 3$, and $2 \to 3$. All the individuals are in state 1 at the time origin, and they are supposed to reach the final absorbing state 3 (typically death) at some future time; along the process, they may experience or not an intermediate event (transient state 2). When the intermediate event represents complications or recurrence during the follow-up, the time spent in state 1 is usually referred to as the disease-free survival time.

Two sets of transition probabilities are to be estimated: for $0 \leq s < t$, $\{\mathsf{P}_{1j}(s,t), j = 1,2,3\}$ and, for $0 < s < t$, $\{\mathsf{P}_{2j}(s,t), j = 2,3\}$. In the case $s = 0$, the transition probabilities $\mathsf{P}_{1j}(0,t)$, $j = 1,2,3$, report the so-called occupation probabilities. It is assumed that n independent, maybe censored trajectories corresponding to n individuals are observed. In this section we review two possible nonparametric approaches to estimate the transition probabilities. The first approach is free of the Markov assumption, while the second one exploits the Markov condition to construct more accurate estimators.

2.1. Non-Markov transition probabilities

Following the notation in de Uña-Álvarez and Meira-Machado (2015), we represent the available information as $(Z_i, T_i, \rho_i, \delta_i)$, $i = 1, \ldots, n$, i.i.d. copies of (Z, T, ρ, δ), where Z and T are respectively the observed sojourn time in state 1 and the observed survival time, and ρ and δ are their corresponding censoring indicators (0 for censoring).

de Uña-Álvarez and Meira-Machado (2015) introduced a new nonparametric estimator of the transition probabilities $\mathsf{P}_{hj}(s,t)$ by considering the subset of individuals observed in state h by time s. To be specific, let $\mathcal{S}_1 = \{i : Z_i > s\}$ and $\mathcal{S}_2 = \{i : Z_i \leq s < T_i\}$ be, respec-

tively, the individuals observed in state 1 and in state 2 by time s. Then, under independent right censoring, the Kaplan-Meier estimator applied to $\{(Z_i, \rho_i), i \in \mathcal{S}_1\}$ is a consistent estimator for $\mathsf{P}_{11}(s,t)$, while the Kaplan-Meier estimators computed from $\{(T_i, \delta_i), i \in \mathcal{S}_1\}$ or $\{(T_i, \delta_i), i \in \mathcal{S}_2\}$ are consistent for $\mathsf{P}_{13}(s,t)$ or $\mathsf{P}_{23}(s,t)$ respectively. For $\mathsf{P}_{12}(s,t)$ and $\mathsf{P}_{22}(s,t)$ the relationships $\mathsf{P}_{12}(s,t) = 1 - \mathsf{P}_{11}(s,t) - \mathsf{P}_{13}(s,t)$ and $\mathsf{P}_{22}(s,t) = 1 - \mathsf{P}_{23}(s,t)$ are used to derive suitable estimators. As mentioned, this approach follows the idea discussed in the seminal paper by Pepe (1991), see also Pepe *et al.* (1991), and it leads to estimators similar to those independently introduced by Titman (2015).

One important property of the new nonparametric estimators in de Uña-Álvarez and Meira-Machado (2015) is that, unlike the Aalen-Johansen estimator (Aalen and Johansen 1978), they are consistent regardless the Markov condition. Compared to other nonparametric estimators with such property (Meira-Machado *et al.* 2006; de Uña-Álvarez 2010), the new estimators are preferable due to their greater accuracy. Indeed, de Uña-Álvarez and Meira-Machado (2015) found through simulations that the new method reports smaller biases and variances. Interestingly, it also avoids the systematic bias of previous proposals (Meira-Machado *et al.* 2006; Allignol, Beyersmann, Gerds, and Latouche 2014) when the support of the censoring time is strictly contained in that of the survival time, which often occurs in practice due to an insufficient follow-up time.

In practice, standard errors and confidence intervals are often demanded. In de Uña-Álvarez and Meira-Machado (2015) the simple bootstrap was suggested to this end. The simple bootstrap just generates B samples of size n from the data, by sampling with replacement each datum with equal probability $1/n$. Then, the variance is estimated by the sampling variance of the estimator when computed along the B bootstrap resamples. An alternative approach to estimate the sampling variance is through plug-in methods, which proceed by replacing the asymptotic variance of the estimator by an empirical counterpart. Explicit formulae for the asymptotic variances of de Uña-Álvarez and Meira-Machado (2015)'s estimators were provided in the web appendices of that paper. From these expressions, plug-in estimators can be introduced in a straightforward way. For example, for $\mathsf{P}_{12}(s,t)$ the asymptotic variance equals

$$\sigma_{12}^{(s)}(t) = \mathsf{E}\left\{[\psi_t^{(s)}(Z_1, \rho_1) - \xi_t^{(s)}(T_1, \delta_1)]^2 I(Z_1 > s)\right\} / \mathsf{P}(Z_1 > s)^2, \tag{3}$$

where the transformations $\psi_t^{(s)}$ and $\xi_t^{(s)}$ can be estimated from the data (see Appendix A). The expectation and the probability in (3) can be replaced by sampling averages to construct the final estimator. One can proceed similarly for the other transition probabilities to derive their plug-in variance estimators. Indeed, for $\mathsf{P}_{11}(s,t)$, $\mathsf{P}_{13}(s,t)$, $\mathsf{P}_{22}(s,t)$ and $\mathsf{P}_{23}(s,t)$, the estimators reduce to Greenwood-type formulae when applied to the specific subsets $\mathcal{S}_1$ and $\mathcal{S}_2$. Details are given in Appendix A. The plug-in approach is less computationally demanding than the bootstrap approach, and therefore it is implemented in the **TP.idm** package.

2.2. Aalen-Johansen transition probabilities

For any s, t with $0 \leq s < t$, for Markov models we have

$$\mathsf{P}_{hj}(s,t) = \mathsf{P}\left(X(t) = j | X(s) = h, H_{s^-}\right) = \mathsf{P}\left(X(t) = j | X(s) = h\right). \tag{4}$$

Thus the future of the process after time s depends only on the state occupied by that time. This is an important class of models where efficient estimators of transition probabilities can be introduced.

Andersen *et al.* (1993) defined the integrated hazard matrix $\mathbf{A} = (A_{hj})$ where $A_{hj}(t) = \int_0^t \alpha_{hj}(s)ds$ for all h, j with $\alpha_{hh}(t) = -\sum_{h \neq j} \alpha_{hj}(t)$. Then, the transition probability matrix $\mathsf{P}(s,t) = (\mathsf{P}_{hj}(s,t))$ is given as the product-integral

$$\mathsf{P}(s,t) = \prod_{(s,t]} (I + d\mathbf{A}(u)), \tag{5}$$

where $\mathbf{A} = (A_{hj})$. The Nelson-Aalen estimator of A_{hj}, denoted by $\hat{A}_{hj}$, is defined as

$$\hat{A}_{hj}(t) = \begin{cases} \int_0^t J_h(u)(Y_h(u))^{-1} dN_{hj}(u), & h \neq j, \\ -\sum_{h \neq j} \hat{A}_{hj}(t), & h = j, \end{cases} \tag{6}$$

where $J_h(u) = I(Y_h(u) > 0)$. Here, $Y_h(u)$ and $N_{hj}(u)$ denote, respectively, the number of individuals observed in state h just prior time u, and the number of observed direct transitions from h to j in the time interval $[0, u]$. Then, one can estimate the transition probability matrix (5) by the $N \times N$ matrix

$$\hat{\mathsf{P}}(s,t) = \prod_{(s,t]} \left(I + d\hat{\mathbf{A}}(u)\right), \tag{7}$$

where $\hat{\mathbf{A}} = (\hat{A}_{hj})$ is the matrix with entries the elements in (6). This is the so-called Aalen-Johansen estimator (Aalen and Johansen 1978).

It has been shown that the Aalen-Johansen estimator may be unsuitable when the process does not fulfill the Markov condition (4), see Meira-Machado *et al.* (2006). The Markov condition is violated when, e.g., the risk of death increases shortly after the recurrence of a disease; in such a case, the length of stay in the intermediate state is relevant for prognosis, thus invalidating the memoryless property of Markov processes. In Section 4 we compare the Aalen-Johansen estimator to the Markov-free nonparametric estimator introduced in Section 2.1 in a practical setting, to show the systematic biases that may appear in non-Markov scenarios.

The variance of the Aalen-Johansen estimator can be calculated by a Greenwood-type formula. Specifically, the covariance matrix $\hat{\mathsf{P}}(s,t)$ is given by (cfr. Andersen *et al.* 1993, Equation 4.4.17)

$$\widehat{\mathsf{COV}}\left(\hat{\mathsf{P}}(s,t)\right) = \int_s^t \hat{\mathsf{P}}(u,t)^\top \otimes \hat{\mathsf{P}}(s,u-) \widehat{\mathsf{COV}}\left(d\hat{\mathbf{A}}(u)\right) \hat{\mathsf{P}}(u,t) \otimes \hat{\mathsf{P}}(s,u-)^\top, \tag{8}$$

where $\widehat{\mathsf{COV}}\left(d\hat{\mathbf{A}}\right)$ is the covariance of the matrix $d\hat{\mathbf{A}}$. This expression can be greatly simplified through a recursion formula (same reference).

2.3. Confidence intervals

Let $\hat{\mathsf{P}}_{hj}(s,t)$ be the transition probability from state h to state j between times s and t, estimated by the non-Markovian estimator (Section 2.1) or by the Aalen-Johansen estimator (7). Let $\hat{\sigma}_{hj}(s,t)$ be the empirical standard error (bootstrap-based or Greenwood-based (8), as discussed above), and let $z_{\alpha/2}$ be the upper $\alpha/2$ quantile of the standard normal distribution.

The linear confidence interval for $\hat{\mathsf{P}}_{hj}(s,t)$ is defined as

$$\hat{\mathsf{P}}_{hj}(s,t) \pm z_{\alpha/2} \cdot \hat{\sigma}_{hj}(s,t). \tag{9}$$

In addition to the linear confidence interval it is possible to consider transformations to improve the confidence intervals in the case of small sample size such as log transformation (10), log-log transformation (11) and complementary log-log transformation (12), see Thomas and Grunkemeier (1975) and Kalbfleisch and Prentice (2002):

$$\hat{\mathsf{P}}_{hj}(s,t)\exp\left\{\frac{\pm z_{\alpha/2}\cdot\hat{\sigma}_{hj}(s,t)}{\hat{\mathsf{P}}_{hj}(s,t)}\right\}, \tag{10}$$

$$\hat{\mathsf{P}}_{hj}(s,t)\exp\left\{\frac{\pm z_{\alpha/2}\cdot\hat{\sigma}_{hj}(s,t)}{\hat{\mathsf{P}}_{hj}(s,t)\log(\hat{\mathsf{P}}_{hj}(s,t))}\right\}, \tag{11}$$

$$1-\left(1-\hat{\mathsf{P}}_{hj}(s,t)\right)\exp\left\{\frac{\pm z_{\alpha/2}\cdot\hat{\sigma}_{hj}(s,t)}{\left(1-\hat{\mathsf{P}}_{hj}(s,t)\right)\log\left(1-\hat{\mathsf{P}}_{hj}(s,t)\right)}\right\}. \tag{12}$$

These four methods are available in the **TP.idm** package.

3. The TP.idm package

The package **TP.idm** was developed to calculate estimates for the transition probability matrix (if $s > 0$) or state occupation probability matrix (if $s = 0$) in the illness-death model. This package includes the novel non-Markovian estimator (de Uña-Álvarez and Meira-Machado 2015) described in Section 2.1, and the Aalen-Johansen estimator for a Markov model, see Section 2.2, Equation 7. Confidence limits and plots are available too.

The main function of the package, `TPidm`, calculates the state occupation probabilities $\mathsf{P}_{hj}(0,t)$, and the transition probabilities $\mathsf{P}_{hj}(s,t)$ for a given time s, estimated by the non-Markovian method or by the Aalen-Johansen approach. The user can fix a specific "future time" t; otherwise the maximum event time is automatically chosen. The function `TPidm` provides confidence intervals for both methods too.

The data frame to be used in the main function of the package must have one row per individual and must include at least the four variables named `time1`, `event1`, `Stime` and `event`, which correspond to the disease free survival time, disease free survival indicator, time to death or censoring, and death indicator, respectively. These are the variables denoted respectively by Z, ρ, T and δ in Section 2.1.

The arguments of the main function `TPidm` are:

- `data`: A data frame as described above.
- `s`: The current time for the transition probabilities to be computed; `s = 0` reports the occupation probabilities.
- `t`: The future time for the transition probabilities to be computed. Default is `"last"` which means the largest time among the uncensored entry times for the intermediate state and the final absorbing state.
- `cov`: A categorial variable for optional by-group analysis; this variable must be a `factor`.
- `CI`: If `TRUE` (default), confidence intervals are computed.

- `level`: Level of confidence intervals. Default is 0.95 (corresponding to 95%).
- `ci.transformation`: Transformation applied to compute confidence intervals. Possible choices are `"linear"`, `"log"`, `"log-log"` and `"cloglog"`, corresponding to formulae (9)–(12) Default is `"linear"`;
- `method`: The method used to compute the transition probabilities. Possible options are `"AJ"` (Aalen-Johansen) or `"NM"` (non-Markovian). Default is `"NM"`.

Additionally, the **TP.idm** package includes `summary`, `print` and `plot` methods for the object returned by function `TPidm`. The `print` and `summary` methods provide details about the multi-state model, the estimates of $\mathsf{P}(s,t)$, and the confidence limits and variances. The `plot` function provides an automatic graphical display for the obtained results.

The main function `TPidm` saves the estimated transition probabilities, their estimated variances, and the corresponding confidence limits in a list, which can be later used to construct plots other than the default ones.

The **TP.idm** package includes another function `test.nm` which performs a graphical test for the Markov condition. This graphical test is a PP-plot which compares the estimations reported by the Aalen-Johansen transition probabilities to their non-Markov counterparts. Since the estimator for $\mathsf{P}_{11}(s,t)$ obtained from both methods is the same, $\mathsf{P}_{11}(s,t)$ is excluded from this graphical test. Also, since the Aalen-Johansen estimator is consistent in the case $s = 0$ (occupation probabilities) regardless the Markov condition, a warning message is reported by the package ("Markov assumption is not relevant for the estimation of occupation probabilities") in this case.

All these functions and parameters are illustrated in Section 4.

4. Application to real data

To illustrate how to use the functions available in the package **TP.idm**, in this section we consider an application to real data. To this end, we use the data frame `colonTP` which is available in package **TP.idm**. This data frame reports information on 929 patients from a large clinical trial on Duke's stage III colon cancer, who underwent a curative surgery for colon-rectal cancer (Moertel and others 1990). In this study, 468 patients developed recurrence and, among these, 414 died, while 38 patients died without recurrence. We model this data through an illness-death progressive model with initial state "alive without recurrence", intermediate state "recurrence", and final absorbing state "death". Our focus is the estimation of the transition probabilities and state occupation probabilities in this model, for all the patients (Section 4.1) and for the three treatment groups (variable `rx`): Observation (315 patients), Levamisole (310), and Levamisole+5 FU (304 patients) (Section 4.2). Below we show the format of the data:

```
R> library("TP.idm")
R> data("colonTP", package = "TP.idm")
R> colonTP[1:6, 1:5]

  time1 event1 Stime event      rx
1   968      1  1521     1 Lev+5FU
```

```
2  3087      0  3087     0 Lev+5FU
3   542      1   963     1     Obs
4   245      1   293     1 Lev+5FU
5   523      1   659     1     Obs
6   904      1  1767     1 Lev+5FU
```

Note that `time1 < Stime` means that a transition from initial state to intermediate state occurred. If `time1 == Stime` and `event1 == 0`, then the patient remained alive and disease-free up to the end of the follow-up; while if `time1 == Stime` and `event1 == 1`, then a direct transition from the initial state to the final (absorbing) state was observed. The data frame `colonTP` reproduces the information in the `colon` object of the package **survival** (Therneau 2017) but re-organized in a suitable way, so only one row is used for each patient. To help the analysis, the value of `Stime` was increased in 0.5 units for the 7 cases reporting a zero transition time from recurrence to death in the data frame `colon`.

4.1. Overall results

We estimate and plot the occupation probabilities along the first year after surgery for the full set of 929 patients by applying the non-Markovian estimator (default method), by running the following code lines:

```
R> nm01 <- TPidm(colonTP, s = 0, t = 365)
R> nm01

Call:
TPidm(data = colonTP, s = 0, t = 365)

Parameters:
s= 0
t= 365
Method= NM
CI= TRUE
CI transformation= linear
Possible transitions:
[1] "1 1" "1 2" "1 3"

Occupation probabilities at time t:


transition      probs      lower     upper     variance
       1 1 0.75242196 0.72469972 0.7801442 2.000600e-04
       1 2 0.16361679 0.14032419 0.1869094 1.412342e-04
       1 3 0.08396125 0.06611612 0.1018064 8.289786e-05

R> plot(nm01)
```

The numerical results indicate that 75% of the patients are still alive and disease-free one year after surgery (variance in estimation: 0.0002; 95% confidence limits: 0.725–0.780), while

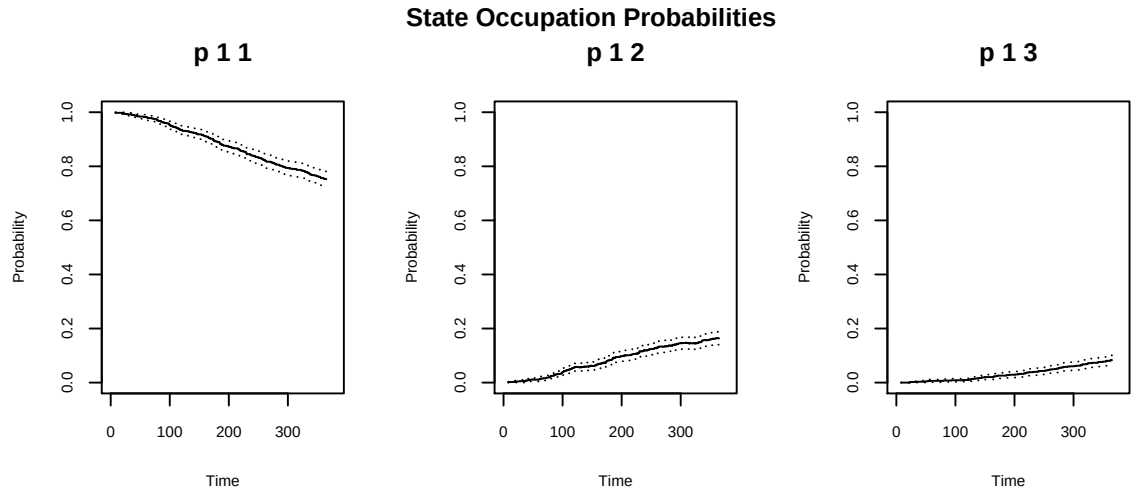

Figure 1: Non-Markovian occupation probabilities along the 365 days after surgery. Colon cancer study.

16% are alive with recurrence by that time. The automatic plot (Figure 1) reports the three occupation probabilities, together with pointwise 95% confidence limits (with the default `ci.transformation = "linear"`), along the first year after surgery, according to the chosen value for the parameter `t = 365`. To obtain a full graphical display along time, one should take the default `t = "last"` instead, see Section 4.2. The (uncensored) entry times for the intermediate and the final states along the interval $[0, 365]$ are saved in the object `nm01$times`; therefore, this object contains all the possible jump points of the empirical transition probabilities along the fixed interval. In this case, `nm01$times` has length 194. It is possible to display estimated occupation probabilities at intermediate times by calling the object `nm01$all.probs[, 1, ]` as follows:

```
R> nm01$all.probs[seq(1, 194, length.out = 5), 1, ]
```

```
     trans
rows        1 1         1 2        1 3
  8   0.9989236 0.001076426 0.00000000
  122 0.9311087 0.058127018 0.01076426
  204 0.8697524 0.100107643 0.03013994
  279 0.8073197 0.136706136 0.05597417
  365 0.7524220 0.163616792 0.08396125
```

This shows for example that, 122 days after surgery ($\approx$ 4 months), the distribution of patients is 93% alive and disease-free, 6% alive with recurrence, and 1% dead. Confidence limits and variances for the intermediate time 122 can be displayed as follows:

```
R> nm01$all.probs[nm01$times == 122, 1:4, ]
```

```
          trans
cols                1 1          1 2          1 3
```

```
  probs    9.311087e-01 5.812702e-02 1.076426e-02
  lower    9.148379e-01 4.313396e-02 4.125039e-03
  upper    9.473795e-01 7.312008e-02 1.740349e-02
  variance 6.891647e-05 5.851734e-05 1.147462e-05
```

We use the following lines to compute the non-Markovian occupation probabilities two years after surgery, and the transition probabilities from time $s = 365$ (one year) to $t = 730$ (two years) too:

```
R> nm02 <- TPidm(colonTP, s = 0, t = 730)
R> nm12 <- TPidm(colonTP, s = 365, t = 730)
R> nm02

Call:
TPidm(data = colonTP, s = 0, t = 730)

Parameters:
s= 0
t= 730
Method= NM
CI= TRUE
CI transformation= linear
Possible transitions:
[1] "1 1" "1 2" "1 3"

Occupation probabilities at time t:


transition     probs     lower     upper     variance
       1 1 0.5994026 0.5679173 0.6308878 0.0002580580
       1 2 0.1744163 0.1511832 0.1976494 0.0001405136
       1 3 0.2261811 0.1992494 0.2531128 0.0001888131

R> nm12

Call:
TPidm(data = colonTP, s = 365, t = 730)

Parameters:
s= 365
t= 730
Method= NM
CI= TRUE
CI transformation= linear
Possible transitions:
[1] "1 1" "1 2" "1 3" "2 2" "2 3"
```

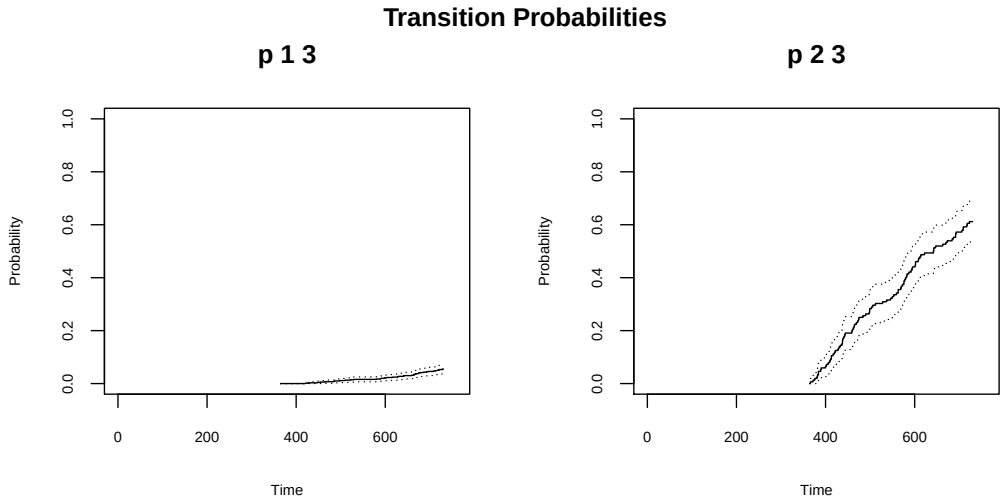

Figure 2: Non-Markovian transition probabilities $\mathsf{P}_{13}(365, t)$ and $\mathsf{P}_{23}(365, t)$ for $t \in (365, 730]$. Colon cancer study.

```
Transition probabilities at time t:


transition     probs      lower      upper     variance
       1 1 0.7966309 0.76680585 0.82645591 2.315611e-04
       1 2 0.1475010 0.12166865 0.17333340 1.737131e-04
       1 3 0.0558681 0.03881836 0.07291783 7.567264e-05
       2 2 0.3881579 0.31125428 0.46506151 1.539563e-03
       2 3 0.6118421 0.53493849 0.68874572 1.539563e-03
```

Note that, when $s > 0$, five estimators are reported, corresponding to the five possible transitions. When $s = 0$, transition probabilities from the intermediate state 2 are not displayed since no individual occupies that state at the time origin. By comparing the reported estimators, one can see that recurrence has a negative impact in the prognosis; indeed, the two-year survival decreases from 94% ($= 100(1 - 0.0559)\%$; confidence limits: 0.9271–0.9612) to 39% (0.3113–0.4651) when one moves from the individuals alive and disease-free one year after surgery to those with recurrence by that time. A graphical comparison of the transition probabilities to the death state for both groups is reported in Figure 2, which is obtained by running the following line:

```
R> plot(nm12, chosen.tr = c("1 3", "2 3"))
```

4.2. By-treatment analysis

We compute the non-Markovian (default method) state occupation probabilities $\mathsf{P}_{1j}(0, t)$, $j = 1, 2, 3$ with `t = "last"` (default future time) for the three treatment groups as follows:

```
R> nm0t_rx <- TPidm(colonTP, s = 0, cov = "rx")
```

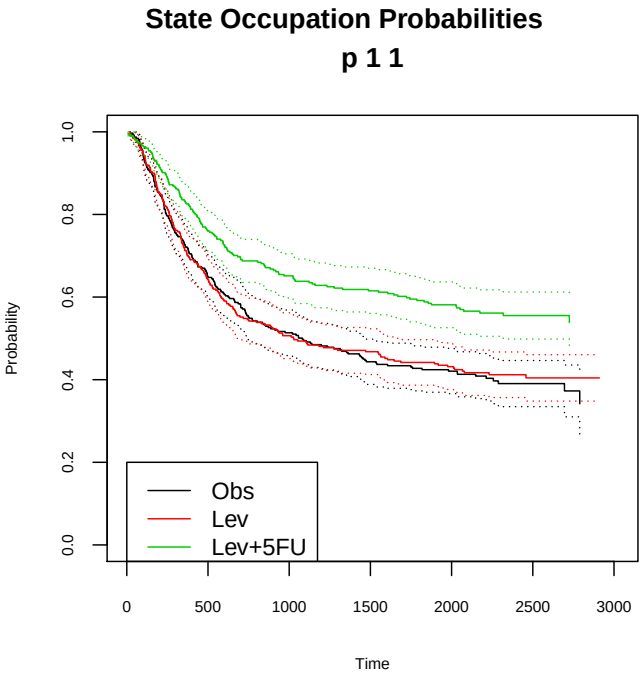

Figure 3: Non-Markovian disease-free survival function $\mathsf{P}_{11}(0,t)$ for the three treatment groups. Colon cancer study.

The by-treatment numerical results can be displayed as before by using the following code line:

```
R> nm0t_rx
```

The disease-free survival function $\mathsf{P}_{11}(0,t)$, together with the corresponding 95% confidence limits (with default `ci.transformation = "linear"`), can be displayed in a single plot by using the following lines:

```
R> plot(nm0t_rx, chosen.tr = c("1 1"), col = 1:3)
R> legend(0, 0.2, legend = c("Obs", "Lev", "Lev+5FU"), lty = 1, col = 1:3)
```

The result is shown in Figure 3. The plot corresponding to $\mathsf{P}_{13}(0,t)$ is given in Figure 4, and it is simply obtained using:

```
R> plot(nm0t_rx, chosen.tr = c("1 3"), col = 1:3)
R> legend(0, 1, legend = c("Obs", "Lev", "Lev+5FU"), lty = 1, col = 1:3)
```

In Figures 3 and 4 it is seen how the combined treatment Levamisole+5 FU improves the disease-free and overall survival functions, as previously reported for this dataset (Moertel and others 1995).

The package **TP.idm** also allows for the computation of the Aalen-Johansen estimator. When one is confident of the Markov assumption, the Aalen-Johansen is preferred over the non-Markovian estimator since it reports a smaller variance in estimation. However, it has been shown that the Aalen-Johansen estimator may be inconsistent when the process does not

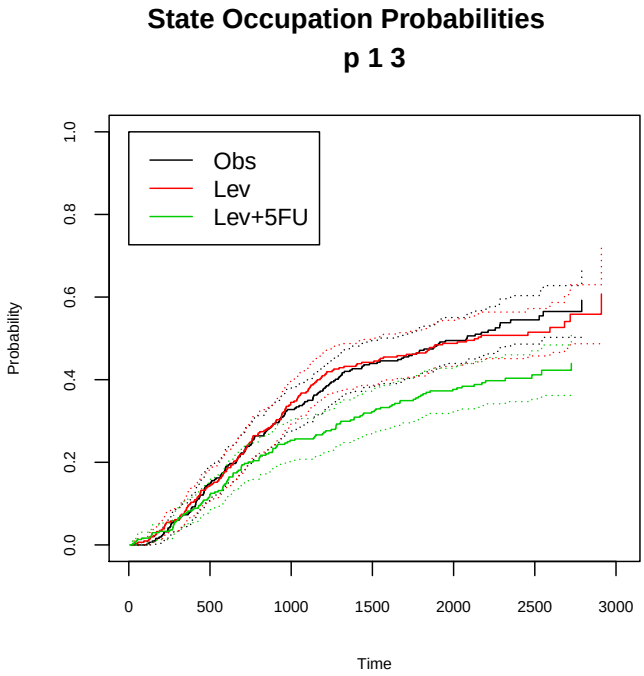

Figure 4: Non-Markovian estimator of $\mathsf{P}_{13}(0,t)$ for the three treatment groups. Colon cancer study.

fulfill the Markov condition (Meira-Machado *et al.* 2006). In the following line we perform a graphical test for the Markov condition in the Observation group:

```
R> test.nm(colonTP[colonTP$rx == "Obs", ], s = 365)
```

Specifically, the plot compares the Aalen-Johansen estimator and the non-Markovian estimator for $\mathsf{P}_{12}(s,t)$, $\mathsf{P}_{13}(s,t)$ and $\mathsf{P}_{22}(s,t)$, for the Observation group and $s = 365$ (Figure 5). Since there exists a deviation of the plots with respect to the straight line $y = x$, one gets some evidence on the lack of Markovianity of the underlying process beyond one year after surgery. Indeed, the test for Markovianity based on the Cox model reported a p value of 0.062 (regression coefficient: -0.000528) for the Observation group, which can be seen by using the function `coxph` of the package **survival**:

```
R> colonTP$entrytime <- colonTP$time1
R> coxph(Surv(time1, Stime, event) ~ entrytime,
+    data = colonTP[colonTP$time1 < colonTP$Stime & colonTP$rx == "Obs", ])
```

Thus, in principle the application of the Aalen-Johansen method is not recommended here, due to possible biases. For further illustration, in Figure 6 we jointly display the non-Markovian estimator and the Aalen-Johansen estimator for $\mathsf{P}_{22}(s,t)$, Observation group and $s = 365$. In this plot the differences between both estimators are clearly seen. The following lines can be used to construct Figure 6:

```
R> plot(TPidm(colonTP[colonTP$rx == "Obs",], s = 365), chosen.tr = c("2 2"))
R> aj1t.Obs <- TPidm(colonTP[colonTP$rx == "Obs",], s = 365, method = "AJ")
R> lines(aj1t.Obs$times, aj1t.Obs$all.probs[ , 1, 4], type = "s", col = 2)
```

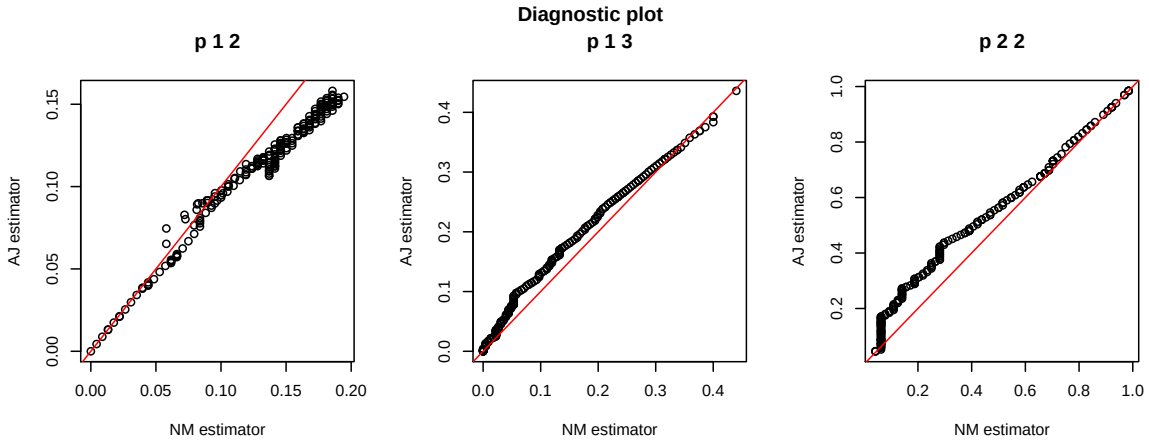

Figure 5: Graphical test for the Markov condition, $s = 365$. Colon cancer study.

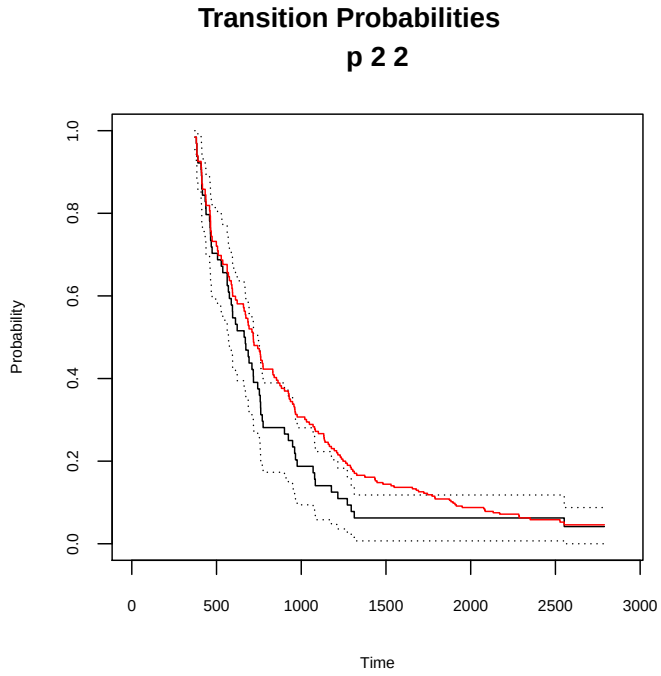

Figure 6: Non-Markovian estimator with 95% pointwise confidence limits (black lines) and Aalen-Johansen estimator (red line) for the transition probability $\mathsf{P}_{22}(s,t)$ for the Observation group and $s = 365$. Colon cancer study.

5. Discussion

Multi-state models are often used to analyze time-to-event data. In the last years, a number of R packages implementing multi-state models techniques have appeared, helping to the dissemination and application of multi-state models in biomedical research, among other fields. The progressive illness-death model is a three-states model with plenty of applications, for which novel statistical methods have been recently proposed. In particular, a lot of emphasis has been put on the nonparametric estimation of the transition probability matrix

for the illness-death model. Alternatives to the classical approach introduced by Aalen and Johansen (1978) include semiparametric approaches (Moreira, de Uña-Álvarez, and Meira-Machado 2013), which allow for a variance reduction; and Markov-free estimators (Allignol *et al.* 2014), with general validity regardless of the Markov condition.

In this paper we have described the **TP.idm** package which implements, for the first time, a novel non-Markovian transition probability matrix for the illness-death model (de Uña-Álvarez and Meira-Machado 2015). The package allows for right censored data, and it provides variance estimates as well as confidence limits. The new method is recommended over previously existing non-Markovian estimators, due to its relatively greater accuracy (same reference). It is also preferred to the Aalen-Johansen estimator when the process under investigation violates the Markov condition since, in such a case, the latter estimator may be systematically biased. Since the Aalen-Johansen estimator is consistent for the estimation of occupation probabilities even in non-Markov scenarios (Datta and Satten 2001), and because of its generally smaller variance, it has been implemented in the **TP.idm** package too. We have compared the computation time of the `TPidm` function to that of the **etm** package, which can be used to obtain the Aalen-Johansen estimator too. The relative performance of these two packages varies depending on the situation (sample size, presence of grouping or rounding in the data) but, generally speaking, they report similar computation times.

Acknowledgment

Work supported by the Grant MTM2014-55966-P of the Spanish Ministerio de Economía y Competitividad.

References

Aalen OO, Johansen S (1978). "An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations." *Scandinavian Journal of Statistics*, **5**(3), 141–150.

Allignol A, Beyersmann J, Gerds T, Latouche A (2014). "A Competing Risks Approach for Nonparametric Estimation of Transition Probabilities in an Non-Markov Illness-Death Model." *Lifetime Data Analysis*, **20**(4), 495–513. doi:10.1007/s10985-013-9269-1.

Allignol A, Beyersmann J, Schumacher M (2008). "**mvna**: An R Package for the Nelson-Aalen Estimator in Multistate Models." *R News*, **8**(2), 48–50.

Allignol A, Schumacher M, Beyersmann J (2011). "Empirical Transition Matrix of Multi-State Models: The **etm** Package." *Journal of Statistical Software*, **38**(4), 1–15. doi:10.18637/jss.v038.i04.

Andersen PK, Borgan Ø, Gill RD, Keiding N (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York. doi:10.1007/978-1-4612-4348-9.

Araújo A, Meira-Machado L, Roca-Pardiñas J (2014). "**TPmsm**: Estimation of the Transition Probabilities in 3-State Models." *Journal of Statistical Software*, **62**(4), 1–29. doi:10.18637/jss.v062.i04.

Balboa-Barreiro V, de Una-Alvarez J, Meira-Machado L (2016). ***TP.idm**: Estimation of Transition Probabilities for the Illness-Death Model*. R package version 1.2, URL https://CRAN.R-project.org/package=TP.idm.

Commenges D (1999). "Multi-State Models in Epidemiology." *Lifetime Data Analysis*, **5**(4), 315–327. doi:10.1023/a:1009636125294.

Datta S, Satten GA (2001). "Validity of the Aalen-Johansen Estimators of Stage Occupation Probabilities and Nelson-Aalen Estimators of Integrated Transition Hazards for Non-Markov Models." *Statistics & Probability Letters*, **55**(4), 403–411. doi:10.1016/s0167-7152(01)00155-9.

de Uña-Álvarez J (2010). "Recent Developments in Censored, Non-Markov Multi-State Models." In C Borgelt, G González Rodriguez, W Trutschnig, M Asunción Lubiano, M Ángeles Gil, P Grzegorzewski, O Hryniewicz (eds.), *Combining Soft Computing and Statistical Methods in Data Analysis*, volume 77 of *Advances in Intelligent and Soft Computing*, pp. 173–179. Springer-Verlag. doi:10.1007/978-3-642-14746-3_22.

de Uña-Álvarez J, Meira-Machado L (2015). "Nonparametric Estimation of Transition Probabilities in the Non-Markov Illness-Death Model: A Comparative Study." *Biometrics*, **71**(2), 364–375. doi:10.1111/biom.12288.

de Wreede LC, Fiocco M, Putter H (2011). "**mstate**: An R Package for the Analysis of Competing Risks and Multi-State Models." *Journal of Statistical Software*, **38**(7), 1–30. doi:10.18637/jss.v038.i07.

Ferguson N, Datta S, Brock G (2012). "**msSurv**: An R Package for Nonparametric Estimation of Multistate Models." *Journal of Statistical Software*, **50**(14), 1–24. doi:10.18637/jss.v050.i14.

Hougaard P (1999). "Multi-State Models: A Review." *Lifetime Data Analysis*, **5**(3), 239–264. doi:10.1023/a:1009672031531.

Hougaard P (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag. doi:10.1007/978-1-4612-1304-8.

Jackson CH (2011). "Multi-State Models for Panel Data: The **msm** Package for R." *Journal of Statistical Software*, **38**(8), 1–29. doi:10.18637/jss.v038.i08.

Kalbfleisch JD, Prentice RL (2002). *Statistical Analysis of Failure Time Data*. 2nd edition. John Wiley & Sons. doi:10.1002/9781118032985.

Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C (2006). "Nonparametric Estimation of Transition Probabilities in a Non-Markov Illness–Death Model." *Lifetime Data Analysis*, **12**(3), 325–344. doi:10.1007/s10985-006-9009-x.

Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C, Andersen PK (2009). "Multi-State Models for the Analysis of Time to Event Data." *Statistical Methods in Medical Research*, **18**(2), 195–222. doi:10.1177/0962280208092301.

Meira-Machado L, Pardiñas JR (2011). "**p3state.msm**: Analyzing Survival Data from an Illness-Death Model." *Journal of Statistical Software*, **38**(3), 1–18. doi:10.18637/jss.v038.i03.

Moertel CG, others (1990). "Levamisole and Fluorouracil for Adjuvant Therapy of Resected Colon Carcinoma." *New England Journal of Medicine*, **322**(6), 352–358. doi:10.1056/NEJM199002083220602.

Moertel CG, others (1995). "Fluorouracil Plus Levamisole as Effective Adjuvant Therapy after Resection of Stage III Colon Carcinoma: A Final Report." *The Annals of Internal Medicine*, **122**(5), 321–326. doi:10.7326/0003-4819-122-5-199503010-00001.

Moreira AC, de Uña-Álvarez J, Meira-Machado L (2013). "Presmoothing the Aalen-Johansen Estimator in the Illness-Death Model." *Electronic Journal of Statistics*, **7**, 1491–1516. doi:10.1214/13-ejs816.

Pepe MS (1991). "Inference for Events with Dependent Risks in Multiple End-Point Studies." *Journal of the American Statistical Association*, **86**(415), 364–375. doi:10.1080/01621459.1991.10475108.

Pepe MS, Longton G, Thornquist M (1991). "A Qualifier Q for the Survival Function to Describe the Prevalence of a Transient Condition." *Statistics in Medicine*, **10**(3), 413–421. doi:10.1002/sim.4780100313.

Therneau TM (2017). ***survival**: Survival Analysis*. R package version 2.41-3, URL https://CRAN.R-project.org/package=survival.

Thomas DR, Grunkemeier GL (1975). "Confidence Interval Estimation of Survival Probabilities for Censored Data." *Journal of the American Statistical Association*, **70**(352), 865–871. doi:10.1080/01621459.1975.10480315.

Titman AC (2015). "Transition Probability Estimates for Non-Markov Multi-State Models." *Biometrics*, **71**(4), 1034–1041. doi:10.1111/biom.12349.

A. Appendix

In this section we provide the definition of the plug-in variance estimators referred in Section 2.1. To this end, we recall some of the asymptotic results in the Web Appendices of de Uña-Álvarez and Meira-Machado (2015). Specifically, the non-Markovian estimator of $\mathsf{P}_{13}(s,t)$, $\hat{\mathsf{P}}_{13}^{NM}(s,t)$ say, is asymptotically Gaussian with limit variance

$$\sigma_{13}^{(s)}(t) = (1-\mathsf{P}_{13}(s,t))^2 \int_s^t \frac{\mathsf{P}_{13}(s,dx)}{(1-\mathsf{P}_{13}(s,x))S_T^{(s)}(x)S_Z(s)}, \tag{13}$$

where $S_T^{(s)}$ stands for the conditional survival functions of T given $Z>s$, and S_Z denotes the survival function of Z. Note that $S_Z(s)$ can be estimated by n^{-1} times the cardinal of the subset $\mathcal{S}_1$, $\hat{S}_Z(s)=n_{1s}/n$ say, while $S_T^{(s)}(\cdot)$ can be estimated by the empirical survival function of the T_i's computed from the subset $\mathcal{S}_1$, that is, $\hat{S}_T^{(s)}(x) = n_{1s}^{-1}\sum_{i=1}^n I(T_i>x)I(Z_i>s)$. Finally, replace $\mathsf{P}_{13}(s,t)$ in (13) by $\hat{\mathsf{P}}_{13}^{NM}(s,t)$ to get

$$\hat{\sigma}_{13}^{(s)}(t) = n(1-\hat{\mathsf{P}}_{13}^{NM}(s,t))^2 \sum_{i=1}^n \frac{I(T_i\le t)\delta_i I(Z_i>s)}{(\sum_{j=1}^n I(T_j\ge T_i)I(Z_j>s))^2}, \tag{14}$$

a Greenwood-type formula applied to the subsample $\mathcal{S}_1$. This leads to the plug-in variance estimator $\widehat{\mathsf{VAR}}(\hat{\mathsf{P}}_{13}^{NM}(s,t)) = \hat{\sigma}_{13}^{(s)}(t)/n$.

Similarly, we obtain $\widehat{\mathsf{VAR}}(\hat{\mathsf{P}}_{22}^{NM}(s,t)) = \widehat{\mathsf{VAR}}(\hat{\mathsf{P}}_{23}^{NM}(s,t)) = \hat{\sigma}_{22}^{(s)}(t)/n$, where

$$\hat{\sigma}_{22}^{(s)}(t) = n\hat{\mathsf{P}}_{22}^{NM}(s,t)^2 \sum_{i=1}^n \frac{I(T_i\le t)\delta_i I(Z_i\le s<T_i)}{(\sum_{j=1}^n I(T_j\ge T_i)I(Z_j\le s<T_j))^2} \tag{15}$$

is a plug-in estimator for the limit variance $\sigma_{22}^{(s)}(t)$ in the Web Appendices of de Uña-Álvarez and Meira-Machado (2015). Again, Equation 15 defines a Greenwood-type estimator, computed in this case from the subset $\mathcal{S}_2$.

Finally, introduce the transformations

$$\xi_t^{(s)}(T_i,\delta_i) = (1-\mathsf{P}_{13}(s,t))\left\{\frac{I(T_i\le t)\delta_i}{S_T^{(s)}(T_i)} - \int_s^{\min(T_i,t)} \frac{\mathsf{P}_{13}(s,dx)}{(1-\mathsf{P}_{13}(s,x))S_T^{(s)}(x)}\right\} \tag{16}$$

and

$$\psi_t^{(s)}(T_i,\delta_i) = \mathsf{P}_{11}(s,t)\left\{\frac{I(Z_i\le t)\rho_i}{S_Z^{(s)}(Z_i)} + \int_s^{\min(Z_i,t)} \frac{\mathsf{P}_{11}(s,dx)}{\mathsf{P}_{11}(s,x)S_Z^{(s)}(x)}\right\}, \tag{17}$$

where $S_Z^{(s)}$ denotes the conditional survival function of Z given $Z>s$. Then, the asymptotic variance of $\hat{\mathsf{P}}_{12}^{NM}(s,t)$ is given by Equation 3, cfr. de Uña-Álvarez and Meira-Machado (2015, Web Appendices), and it can be estimated by

$$\hat{\sigma}_{12}^{(s)}(t) = [n_{1s}/n]^{-2}\frac{1}{n}\sum_{i=1}^n [\hat{\psi}_t^{(s)}(Z_i,\rho_i) - \hat{\xi}_t^{(s)}(T_i,\delta_i)]^2 I(Z_i>s). \tag{18}$$

Here, $\hat{\xi}_t^{(s)}$ and $\hat{\psi}_t^{(s)}$ stand for the natural estimators of the transformations (16) and (17), which are obtained when replacing the transition probabilities by their non-Markovian estimators,

and the conditional survival functions $S_T^{(s)}$ and $S_Z^{(s)}$ by their empirical counterparts, namely $\hat{S}_T^{(s)}(x) = n_{1s}^{-1} \sum_{i=1}^{n} I(T_i \geq x) I(Z_i > s)$ and $\hat{S}_Z^{(s)}(x) = n_{1s}^{-1} \sum_{i=1}^{n} I(Z_i \geq x) I(Z_i > s)$. The variance of $\hat{\mathsf{P}}_{12}^{NM}(s,t)$ is then estimated by $\widehat{\mathsf{VAR}}(\hat{\mathsf{P}}_{12}^{NM}(s,t)) = \hat{\sigma}_{12}^{(s)}(t)/n$.

Affiliation:

Jacobo de Uña-Álvarez
Department of Statistics and Operations Research & CINBIO
Faculty of Economics and Business
Universidade de Vigo
36310 Vigo, Spain
E-mail: jacobo@uvigo.es
URL: http://jacobo.webs.uvigo.es

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd