

msSurv, an R Package for Nonparametric Estimation of Multistate Models

Nicole Ferguson
University of Louisville

Somnath Datta
University of Louisville

Guy Brock
University of Louisville

Abstract

We present an R package, **msSurv**, to calculate the marginal (that is, not conditional on any covariates) state occupation probabilities, the state entry and exit time distributions, and the marginal integrated transition hazard for a general, possibly non-Markov, multistate system under left-truncation and right censoring. For a Markov model, **msSurv** also calculates and returns the transition probability matrix between any two states. Dependent censoring is handled via modeling the censoring hazard through observable covariates. Pointwise confidence intervals for the above mentioned quantities are obtained and returned for independent censoring from closed-form variance estimators and for dependent censoring using the bootstrap. Note that this vignette is adapted from the publication on **msSurv** in the *Journal of Statistical Software* (Ferguson *et al.* 2012).

Keywords: survival, R package, nonparametric estimation, multistate models.

1. Introduction

Multistate models are systems of multivariate survival data where individuals transition through a series of distinct states following certain paths of possible transitions. These systems are illustrated by a directed graph, where distinct states are treated as nodes and possible transitions are considered directed edges. Transitions between states may be reversible or irreversible while states can be either absorbing or transient. Multistate models have a wide range of application including epidemiology, dentistry, clinical trials, reliability studies in engineering, and medicine where individuals progress through the different states of a disease such as cancer and AIDS. Data in these applications are often subject to right censoring and possibly left truncation.

Aalen (1978, see also Nelson 1972) proposed an estimator for the integrated hazard under a broad class of counting process models. Aalen and Johansen (1978) obtained an estimator for the transition probability matrix and subsequently state occupation probabilities through product limit integration of the Nelson-Aalen estimator. Datta and Satten (2001) established that the resulting estimators of state occupation probabilities remained valid even when the process is non-Markovian. Datta and Satten (2002) also proposed an estimator for state occupation probabilities that can handle state dependent censoring and other flexible models through a weighting function based on the censoring scheme. Estimation of state entry and exit distribution functions are also of interest (Pepe 1991; Datta and Ferguson 2011), and can be calculated through normalized sums of state occupation probabilities.

Several R (R Development Core Team 2012) packages are available on the Comprehensive R Archive Network (CRAN, <http://www.r-project.org>) for use with multistate models. Recently, *The Journal of Statistical Software* published a special volume on *Competing Risks and Multi-State Models* featuring papers on the **msm** (Jackson 2011), **mstate** (de Wreede *et al.* 2011), **etm** (Allignol *et al.* 2011), and **p3state.msm** (Meira-Machado and Roca-Pardinas 2011) packages. Other packages currently available include **changeLOS** (Wrangler *et al.* 2006) and **mvna** (Allignol *et al.* 2008). The **msm** package provides functions for fitting multistate Markov models to panel count data and offers extensions to hidden Markov multistate models and possibly inhomogeneous Markov models. The **mstate** package can be applied to right censored and left truncated data in semiparametric or nonparametric multistate models with or without covariates and it may also be applied to competing risk models. The package offers functions which calculate transition probabilities and standard errors. It also uses Cox regression models to estimate different types of covariate effects. The packages **changeLOS**, **mvna**, and **etm** all provide methods for nonparametric estimation in multistate models. The most specialized package available is **changeLOS**, which is based on methods described in Schulgen and Schumacher (1996) and computes changes in length of hospital stay. It does offer a function to compute the Aalen-Johansen estimator, but it does not provide variance estimates and cannot be applied to left truncated data. The **mvna** package computes the Nelson-Aalen estimator of the cumulative transition hazard for any multistate model with right censored, left truncated data, but does not compute transition probability matrices. The **etm** package calculates the transition probability matrices and corresponding variance estimates for any finite-state multistate model with data subject to right censoring and left truncation. State occupation probabilities may be indirectly found using an initial time of zero in the **etm** package.

No packages currently available estimate state entry and exit time distributions, nor do they estimate desired quantities like transition probabilities for dependent censoring. These missing methods are addressed in the **msSurv** package (Ferguson *et al.* 2012), which we introduce in this paper. **msSurv** calculates nonparametric estimates of marginal quantities (that is, not conditional on any covariates) for time to event multistate data subject to right censoring and possibly left truncation. We assume left truncation occurs with respect to the total time of an individual in the multistate system, so that individuals are not considered for the estimation of transitions prior to their left truncation time. The main function `msSurv()` calculates and returns the marginal state occupation probabilities for a general, possibly non-Markov, multistate system, and provides features not currently found in other R packages. Additionally, state entry and exit time distributions are calculated and returned for states where it is appropriate. The function also calculates and returns the marginal integrated transition hazards and the hazard rate functions, and for a Markov model, the transition probability matrix between any two times. Users specify whether the censoring is independent or state dependent and then appropriate calculations of variance and confidence intervals are performed. Pointwise confidence intervals for the above mentioned quantities are obtained and returned for independent censoring from closed-form variance estimators and from bootstrapping for dependent censoring. The function returns an object of S4 class **msSurv** which includes state and transition information, event times, estimates, variance estimates, confidence intervals, and counting process information. The **msSurv** object can then be summarized or plotted using methods available in the package.

The rest of this paper is organized as follows. Section 2 describes nonparametric estimation

methods which are used in the **msSurv** package. Section 3 describes the implementation of **msSurv** and illustrates available functions through examples. Possible future extensions of our software package are discussed in Section 4.

2. The estimators

Consider a multistate model with a finite state space $\mathcal{S} = \{1, \dots, M\}$ and set of possible transitions between the states in \mathcal{S} . The quantities of interest (like state occupation probabilities and state entry and exit distributions) can be calculated for complete data, when available, and also using estimators obtained from censored data when complete data is not known. It is useful to keep track of all the transitions an individual makes before ending in an absorbing state. Let T_{ik}^* represent the time of the k th transition for individual $i = 1, \dots, n$, where $T_{i0}^* = 0$ and $T_{ik}^* = \infty$ if the i th individual enters the absorbing state before the k th transition is made. Let C_i be the right censoring time for the i th individual, L_i be the left truncation time for the i th individual, and s_{ik} be the state occupied by the i th individual between times T_{ik-1}^* and T_{ik}^* . Let $T_i^* = \sup_k \{T_{ik}^* : T_{ik}^* < \infty\}$ be the time for the last transition for individual i . The collection of all transition times and states occupied by individual i can be denoted as $\mathbf{T}_i^* = (T_{ik}^* : k \geq 1)$ and $\mathbf{s}_i^* = (s_{ik} : k \geq 1)$, respectively. Define an indicator of whether the i th individual was never censored, e.g., $\delta_i = I(C_i > T_i^*)$, and let $T_i = \min(T_i^*, C_i)$.

Usually, one assumes that censoring is independent of the multistate system. However, there are times when censoring and hazards of future transitions are affected by time varying covariates $Z = Z(t)$. In these cases, the covariate variables Z explain the dependence and thus future transitions and censoring events behave conditionally independent given the covariate process Z (Datta and Satten 2001, 2002). Modeling of this censoring hazard using the covariate process is discussed in Section 2.3.

2.1. Nelson-Aalen and Aalen-Johansen estimators

Andersen *et al.* (1995) presented formulas for the Nelson-Aalen estimator for the integrated hazard matrix \mathbf{A} and the Aalen-Johansen estimator of the state occupation probability matrix of a Markov system. The counting process and the number at risk for data subject to left truncation and right censoring are estimated as

$$\widehat{N}_{jj'}(t) = \sum_{i=1}^n \sum_{k \geq 1} I(T_{ik}^* \leq t, C_i \geq T_{ik}^*, L_i < t, s_{ik} = j, s_{ik+1} = j') \quad (1)$$

and

$$\widehat{Y}_j(t) = \sum_{i=1}^n \sum_{k \geq 1} I(T_{ik-1}^* < t \leq T_{ik}^*, C_i \geq t, L_i < t, s_{ik} = j). \quad (2)$$

The Nelson-Aalen estimator of the cumulative hazard is given by

$$\widehat{A}_{jj'}(t) = \begin{cases} \int_0^t \frac{I(\widehat{Y}_j(s) > 0)}{\widehat{Y}_j(s)} d\widehat{N}_{jj'}(s) & j \neq j' \\ -\sum_{j \neq j'} \widehat{A}_{jj'}(t) & j = j', \end{cases} \quad (3)$$

The Aalen-Johansen estimator of the transition probability matrix of a Markov multistate system is obtained by product integration of $\widehat{A}_{jj'}(t)$, i.e.,

$$\widehat{\mathbf{P}}(s, t) = \prod_{(s,t]} \left(I + d\widehat{\mathbf{A}}(u) \right), \quad (4)$$

where $\widehat{\mathbf{A}} = \left\{ \widehat{A}_{jj'} \right\}$, which reduces to simple empirical proportions for the complete data.

A recursive formula for computing the variance of transition probability can be found in Andersen *et al.* (1995) (see formula 4.4.19 on p. 295 for details). The resulting estimator is of the Greenwood-type.

2.2. State occupation probabilities

The state occupation probability answers the marginal question: ‘‘What is the probability that an individual is in state j at time t ?’’ Let $p_j(t) = Pr \{s(t) = j\}$ denote the state occupation probabilities where $s(t)$ is the state occupied by an individual at time t , $j \in \{1, \dots, M\}$. For incomplete data, the state occupation probabilities are estimated as

$$\widehat{p}_j(t) = \sum_{k=1}^M \widehat{p}_k(0) \widehat{p}_{kj}(0, t), \quad (5)$$

where $\widehat{p}_{kj}(0, t)$ is the k, j th element of the matrix $\widehat{\mathbf{P}}(0, t) = \prod_{(0,t)} \left(I + d\widehat{\mathbf{A}}(u) \right)$ and $\widehat{p}_k(0)$ is the initial state occupation proportions for state k . This estimator is in fact the Aalen-Johansen estimator of state occupation and holds its validity regardless of Markovian assumptions (Datta and Satten 2001). Estimation of state occupation probabilities can be extended to dependent censoring and other flexible models by explicitly modeling the censoring process (Datta and Satten 2000, 2002).

2.3. Datta-Satten estimators

Datta and Satten (2002) use the principle of inverse probability of censoring weights (Robins and Rotnitzky 1992) to extend the Nelson-Aalen and Aalen-Johansen estimators to data subject to dependent censoring. The resulting Datta-Satten estimators use a weighting function, denoted by K , which is based on a fitted model of the censoring hazards using Aalen’s linear hazards model with possibly time-dependent covariates Z (Aalen 1980).

Rewighted estimators for the counting process and the size of the at risk sets of complete data are defined as

$$\widehat{N}_{jj'}(t) = \sum_{i=1}^n \sum_{k \geq 1} I(T_{ik}^* \leq t, C_i \geq T_{ik}^*, L_i < t, s_{ik} = j, s_{ik+1} = j') / K_i(T_{ik}^* -) \quad (6)$$

and

$$\widehat{Y}_j(t) = \sum_{i=1}^n \sum_{k \geq 1} I(T_{ik-1}^* < t \leq T_{ik}^*, C_i \geq t, L_i < t, s_{ik} = j) / K_i(t-) \quad (7)$$

where $\widehat{K}_i(t) = \exp \left\{ -\widehat{\Lambda}_C^i(t | \mathbf{Z}_i(t)) \right\}$, $\mathbf{Z}_i(t) = \{Z_i(s) : 0 \leq s < t\}$, and

$$\lambda_C^i(t | \mathbf{Z}_i(t)) = \lim_{dt \rightarrow 0} \frac{\Pr \{C_i \in [t, dt), \delta_i | T_i \geq t > L_i, \mathbf{Z}_i(t), \mathbf{T}_i^*, \mathbf{s}_i^*\}}{dt}.$$

See [Datta and Satten \(2002\)](#) for a formal argument.

For state dependent right censoring (i.e., when $Z(t) = s(t)$), this is equivalent to estimating the censoring hazard by a state specific Nelson-Aalen estimator of censoring ([Datta and Satten 2002](#)). Currently, the **msSurv** package only implements this type of dependent censoring.

Substituting the formulas for the estimated counting process and the number at risk into Equations 3, 4, and 5 yields the Datta-Satten estimators of integrated transition hazards, transition probabilities for Markov Systems and state occupation probabilities for non-Markov systems under dependent censoring. Variance estimates for the Datta-Satten estimators of transition probabilities are obtained using the bootstrap.

2.4. State entry and exit distributions

An important application of state occupation probabilities is computing the state entry and exit time distribution functions. We assume that the only path between the states before and after the state of interest (e.g., state j) is through state j , and that the system is acyclic with respect to state j (i.e., it is not possible to return to state j after leaving it). Suppose X_j denotes the possibly unobserved indicator of an individual ever entering state j . Suppose F_j and G_j denote the state entry and exit time distribution functions, respectively, for the individuals who ever enter state j (i.e., $X_j = 1$). Let \mathcal{S}^j denote the collection of all states which come after state j in the progressive model. The entry time distribution to state j is estimated by taking the normalized sum of the estimated state occupation probabilities of state j and all the other states that come after j in the system, i.e.,

$$\widehat{F}_j(t) = \frac{\sum_{k \in \mathcal{S}^j \cup \{j\}} \widehat{p}_k(t)}{\sum_{k \in \mathcal{S}^j \cup \{j\}} \widehat{p}_k(\infty)}, \quad (8)$$

where $\widehat{p}_k(\infty) = \lim_{t \rightarrow \infty} \widehat{p}_k(t)$. The denominator in Equation 8 is the probability of an individual ever entering state j , while the numerator gives the unconditional probability of having entered state j by time t . The unconditional probability (e.g., the subdistribution function) may be of interest in itself, and the **msSurv** package has the option to return either the normalized or unnormalized versions.

The exit time distribution from a transient state j is estimated by taking the normalized sum of estimated state occupation probabilities of all states that come after state j in the progressive system, i.e.,

$$\widehat{G}_j(t) = \frac{\sum_{k \in \mathcal{S}^j} \widehat{p}_k(t)}{\sum_{k \in \mathcal{S}^j \cup \{j\}} \widehat{p}_k(\infty)}. \quad (9)$$

The denominator here is the same as in Equation 8, and **msSurv** provides the option of returning either the normalized or unnormalized version of the function.

The variance for the state entry and exit time distributions are obtained using the bootstrap.

2.5. Confidence intervals

Pointwise confidence intervals for estimators of transition probabilities and state occupation probabilities follow methods described in Andersen *et al.* (1995). Let $\hat{p}(s, t)$ be the transition probability between two states in the system (the subscripts are omitted to simplify the notation) between times s and t and let $\hat{\sigma}(s, t)$ be the corresponding variance estimate. Then the linear confidence interval for $\hat{p}(s, t)$ is defined as

$$\hat{p}(s, t) \pm c_{\alpha/2} \hat{\sigma}(s, t),$$

where $c_{(\alpha/2)}$ is the upper $\alpha/2$ percentile of the standard normal distribution. It may be beneficial to consider transformations to improve estimation especially in the case of small sample sizes (Bie *et al.* 1987; Thomas and Grunkemeier 1975). Borgan and Liestol (1990) suggested a log transformation to improve small sample properties. The resulting formula for the confidence interval is

$$\hat{p}(s, t) \exp \left\{ \frac{\pm c_{\alpha/2} \hat{\sigma}(s, t)}{\hat{p}(s, t)} \right\},$$

Other transformations to improve small sample properties include the log-log transformation, proposed in the first edition of Kalbfleisch and Prentice (2002) and defined as

$$\hat{p}(s, t) \exp \left\{ \frac{\pm c_{\alpha/2} \hat{\sigma}(s, t)}{\hat{p}(s, t) \log(\hat{p}(s, t))} \right\},$$

and the complementary log-log transformation

$$1 - (1 - \hat{p}(s, t)) \exp \left\{ \frac{\pm c_{\alpha/2} \hat{\sigma}(s, t)}{(1 - \hat{p}(s, t)) \log(1 - \hat{p}(s, t))} \right\}.$$

Confidence intervals for state occupation probabilities are calculated using $s = 0$ in the formulas above.

Confidence intervals for state entry and exit time distributions for state i at time t are calculated using $\hat{F}_i(t)$ and $\hat{G}_i(t)$ instead of $\hat{p}(s, t)$, with corresponding variance estimate $\hat{\sigma}(t)$ obtained using the bootstrap.

3. msSurv package implementation

The **msSurv** package may be applied to any general multistate model with data subject to right censoring and possibly left truncation. **msSurv** is written entirely in the R programming language using S4 classes and methods, and is available for download from CRAN. It can be installed on all operating systems for which the R software is installed. Other package dependencies include the **graph** package (Gentleman *et al.* 2010) and the **lattice** package (Sarkar 2008).

The main function of the package, **msSurv**, calculates the counting processes and risk sets according to both Andersen *et al.* (1995) and Datta and Satten (2001), as well as the state occupation probabilities and transition probabilities described in Section 2.5. The **msSurv** package contains a function to calculate the transition probabilities between two specific times (**Pst**), a function to display the state occupation probabilities at a specific time t (**SOpt**), a

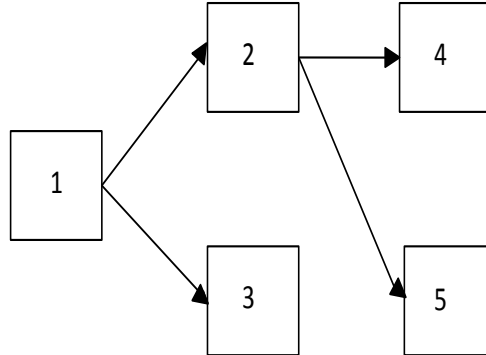


Figure 1: The five state model for the simulated multistate data discussed in Section 3.1.

function to display the state entry and exit time distributions at a specific time t (`EntryExit`), as well as `print`, `plot`, and `summary` methods for `msSurv` objects.

We will illustrate the application of the `msSurv` package through three examples. The first example uses simulated data with independent right censoring, the second example uses a simulated data set with left truncation and independent right censoring, and the third example uses a bone marrow transplant data set from Klein and Moeschberger (2010) with state dependent right censoring.

3.1. A 5 state example

We consider a five-state progressive model with the tree structure illustrated in Figure 1. We simulated a data set of 1000 individuals subject to independent right censoring with 60% of individuals starting in state 1 at time 0 and 40% starting in state 2. Those in state 1 remain there until they transition to the transient state 2 or the terminal state 3. Individuals in state 2 remain there until they transition to either terminal state 4 or 5.

A right censoring time is generated for each of the 1000 individuals using the log normal distribution with log mean -0.5 and log standard deviation 2. For individuals starting in state 1, two times are generated using the Weibull distribution using a sample size of 600 and shape parameter of 2 to reflect transition times between states 1 and 2 and states 1 and 3. Times are compared and the minimum time is kept as the event time and the corresponding state is recorded. Then, two additional times are generated to reflect transitions between states 2 and 4 and states 2 and 5. These times are generated using the formula $T_2 = D^{-1}(D(T_1) + R_2\{1 - D(T_1)\})$ where T_1 is the first transition time, R_2 is a random number generated from $U(0, 1)$ independent of T_1 , $D(\cdot)$ denotes the distribution function for the Weibull distribution with shape parameter 2, $D^{-1}(\cdot)$ denotes the corresponding quantile function. The minimum of these two times and the censoring times are compared and the

minimum time is taken as the event time and the corresponding state is recorded. For individuals starting in state 2, two times are generated using the Weibull distribution with a sample size of 400 and shape parameter of 2 to reflect transitions between states 2 and 4 and states 2 and 5. These times are then compared with the corresponding censoring times and the minimum time is taken as the event time and the state information is recorded. All times were rounded to the fourth decimal place for clarity of presentation. The simulated data is available as `RCdata` in package `msSurv`.

We begin by loading the package.

```
R> library("msSurv")
```

Now we load the data.

```
R> data("RCdata")
```

Data should be in a data frame with column names `"id"`, `"stop"`, `"start.stage"`, and `"end.stage"`, where `"id"` is the individual's identification number, `"stop"` is the transition time from state j to j' , `"start.stage"` is the state the individual is transitioning from (i.e., j), and `"end.stage"` is the state the individual is transitioning to (i.e., j') and equals 0 if right censored.

```
R> RCdata[70:76, ]
```

	id	stop	start.stage	end.stage	
	70	57	0.3086	1	3
	71	58	0.5322	1	2
	72	58	0.6333	2	5
	73	59	0.3330	1	2
	74	59	0.5824	2	5
	75	60	0.7722	1	0
	76	61	0.4096	1	3

Now we specify the tree structure for this multistate system. First, input the states in the multistate system as a list or character vector of the state names. Then, store the transition information as a list of possible states with allowed transitions being lists of edges. For terminal states, the lists of edges will be `NULL`. Nodes correspond to the states in the model and edges refer to the allowed transitions.

```
R> Nodes <- c("1", "2", "3", "4", "5")
R> Edges <- list("1" = list(edges = c("2", "3")),
+              "2" = list(edges = c("4", "5")),
+              "3" = list(edges = NULL),
+              "4" = list(edges = NULL),
+              "5" = list(edges = NULL))
```

The tree structure is then specified using the `graph` package (Gentleman *et al.* 2010) by creating a `graphNEL` object as below with nodes and edges defined above.


```
R> treeobj <- new("graphNEL", nodes = Nodes, edgeL = Edges,
+               edgemode = "directed")
```

Now we will call the `msSurv` function to perform nonparametric estimation for this simulated example.

```
R> ex1 <- msSurv(RCdata, treeobj, bs = TRUE)
```

Entry distributions calculated for states 2 3 4 5 .

Exit distributions calculated for states 1 2 .

Using bootstrap to calculate variances ...

Results of the analysis can be viewed using the `print`, `plot`, and `summary` methods, as well as the `Pst`, `SOpt` and `EntryExit` functions, available for the `msSurv` object `ex1`. The `print` method is discussed in this section while the `summary` and `plot` methods are discussed in Sections 3.2 and 3.3, respectively.

The `print` method gives an overview of the model, specifying the number of transient and absorbing states and identifying the states and possible transitions in the system.

```
R> print(ex1)
```

An object of class 'msSurv'.

Allowable States: 1 2 3 4 5

The multistate model has 2 transient state(s) and 3 absorbing state(s).

Number of timepoints with transitions: 645

Allowable transitions: 1->2, 1->3, 2->4, 2->5

For a listing of available slots and methods see 'class?msSurv'

See also:

'SOpt' for calculation of state occupation probabilities

'Pst' for calculation of transition probabilities and

'EntryExit' for calculation of state entry / exit distributions.

The `msSurv` package contains two functions for the user to easily access estimators of transition and state occupation probabilities at specific time points. The package also contains a function for the user to access state entry and exit time distribution estimators at specific time points.

The transition probabilities between any two times s and t are computed using the `Pst` function. The `Pst` function takes an `msSurv` object, a starting time s , and an ending time t and prints the transition probability matrix $\hat{P}(s, t)$. For example, the following code produces the transition probability matrix $\hat{P}(1, 3.1)$:

```
R> Pst(ex1, s = 1, t = 3.1)
```

```
Estimate of P(1, 3.1)
```

```
  cols
  1 2    3    4    5
1 0 0 0.4805 0.307 0.2124
2 0 0 0.0000 0.586 0.4140
3 0 0 1.0000 0.000 0.0000
4 0 0 0.0000 1.000 0.0000
5 0 0 0.0000 0.000 1.0000
```

The corresponding covariance matrix for this transition probability is printed by adding the argument `covar = TRUE`.

```
R> Pst(ex1, s = 1, t = 3.1, covar = TRUE)
```

State occupation probabilities at a specific time t are given using the `SOPt` function. This function takes a `msSurv` object as well as time t as arguments. The default time t is the maximum event time in the data set. Individuals may start in any state in the system at time 0. The function prints the state occupation probabilities for all states in the system at time t . For example, to find the state occupation probabilities at time $t = 0.85$.

```
R> SOPt(ex1, t = 0.85)
```

The state occupation probabilities at time 0.85 are:

```
State 1: 0.1415
State 2: 0.2179
State 3: 0.2127
State 4: 0.2028
State 5: 0.2252
```

The corresponding variance estimates are provided using the argument `covar = TRUE`. Variance estimates are found by evaluating the formula in Andersen *et al.* (1995) at $\hat{P}(0, t)$ for data where all the individuals start in a single initial state at time 0. The bootstrap is used to find variance estimates of state occupation for data where individuals start in different states of the system.

```
R> SOPt(ex1, t = 0.85, covar = TRUE)
```

The `EntryExit` function in `msSurv` displays the state entry and exit time distributions at a specific time t . This function takes an `msSurv` object and time t as arguments and displays the either normalized (if `norm = TRUE`, the default) or non-normalized (subdistribution) state entry and exit distributions. State entry distributions are only displayed for non-initial states while state exit distributions are only displayed for non-terminal states. Estimates are rounded to four decimal places by default, but the user may specify a different number through the `deci` argument. For example, the normalized state entry and exit distributions at time 1 are displayed by

```
R> EntryExit(ex1, t = 1)
```

```
The normalized state entry distributions at time 1 are:
```

```
State 2 : 0.9587
```

```
State 3 : 0.9018
```

```
State 4 : 0.7172
```

```
State 5 : 0.7794
```

```
Normalized state entry distributions not estimated for state(s) 1
```

```
The normalized state exit distributions at time 1 are:
```

```
State 1 : 0.9427
```

```
State 2 : 0.7467
```

```
Normalized state exit distributions not estimated for state(s) 3 4 5
```

The bootstrap (via the `bs = TRUE` argument in the call to `msSurv`) is needed to obtain variance estimates for the state entry and exit time distributions. If this was done, then the `var = TRUE` argument in the call to the `EntryExit` function can be used to display the variance estimates. Note that all of the function `Pst`, `SOpt`, and `EntryExit` return their results invisibly, so that users can save and access the estimates for later use if desired. For example, in the following code chunk the results of the `EntryExit` call are stored in `res`, which consists of a list with items `entry.norm`, `entry.var.norm`, `exit.norm`, and `exit.var.norm`. If instead subdistributions were requested, (through the argument `norm = FALSE`), then the elements in the list would be the same except with ‘sub’ in place of ‘norm’.

```
R> norm.dist <- EntryExit(ex1, t = 1, var = TRUE)
```

```
R> sub.dist <- EntryExit(ex1, t = 1, norm = FALSE, var = TRUE)
```

```
R> names(norm.dist)
```

```
[1] "entry.norm"      "exit.norm"      "entry.var.norm" "exit.var.norm"
```

```
R> names(sub.dist)
```

```
[1] "entry.sub"      "exit.sub"      "entry.var.sub" "exit.var.sub"
```

3.2. Left truncation and right censoring example

We consider an irreversible three-state illness-death model with data subject to independent right censoring and left truncation (Andersen *et al.* 1995). We simulated a data set of 1000 individuals starting in state 1 at time 0. Individuals remain in state 1 until they transition to the transient state 2 (ill) or the terminal state 3 (death). Individuals in state 2 remain there until they transition to the terminal state 3 (death).

Two times are generated using the Weibull distribution with a sample size of 1000 and shape parameter of 2 to reflect transition times for either illness or death. Right censoring and

left truncation times are generated using the log normal distribution with mean -0.5 and standard deviation 2 on the log scale. We assume 20% of individuals have a left truncation time of 0. Only individuals whose left truncation times were less than the terminal event times or censored event times were kept. The left truncation time was taken to be the time the individual entered the study. Individuals were not included in the at risk set before their left truncation time. Times for these individuals were compared and the minimum time was kept as the event time and the corresponding state was recorded. Then, another time was generated reflecting the transition between states 2 and 3 using the formula $T_2 = D^{-1}(D(T_1) + R_2\{1 - D(T_1)\})$ where T_1 is the first transition time, R_2 is a random number generated from $U(0,1)$ independent of T_1 , $D(\cdot)$ denotes the distribution function for the Weibull distribution with shape parameter 2, $D^{-1}(\cdot)$ denotes the corresponding quantile function. These “death” times are then compared to the censoring times and the minimum time is kept as the event time and the corresponding state information is recorded. All times were rounded to the fourth decimal place for clarity of presentation. The simulated data is available as `LTRCdata` in package `msSurv`.

Begin by loading the data

```
R> data("LTRCdata")
```

Data should be in a data frame with column names "id", "start", "stop", "start.stage", and "end.stage" where "id" is the individual's identification number, "start" is the start time for the period of observation after the individual enters state j (and is the left truncation time for the first observed transition), "stop" is the transition time from state j to j' , "start.stage" is the state the individual is transitioning from (i.e., j), and "end.stage" is the state the individual is transitioning to (and equals 0 if right censored).

```
R> LTRCdata[489:494, ]
```

	id	start	stop	start.stage	end.stage
489	367	0.0696	0.2682	2	3
490	368	0.0000	0.9334	1	3
491	369	0.0000	0.1598	1	3
492	370	0.0000	0.9943	1	2
493	370	0.9943	1.0921	2	3
494	371	0.0000	0.6698	1	3

Now we specify the tree structure.

```
R> Nodes <- c("1", "2", "3")
R> Edges <- list("1" = list(edges = c("2", "3")),
+              "2" = list(edges = c("3")),
+              "3" = list(edges = NULL))
R> LTRCtree <- new("graphNEL", nodes = Nodes, edgeL = Edges,
+               edgemode = "directed")
```

The `msSurv` function is called to perform nonparametric estimation for this simulated example. Since the data is subject to left truncation, we must add the argument `LT = TRUE` to the call of `msSurv`.

```
R> ex2 <- msSurv(LTRCdata, LTRCtree, LT = TRUE)
```

Entry distributions calculated for states 3 .

Exit distributions calculated for states 1 .

When left truncation is present, starting proportions in each state are calculated empirically based on the initial states of those individuals whose left truncation time is prior to the first observed transition time.

Results of the analysis will be illustrated through the `summary` method available for the `msSurv` object `ex2`. The `summary` method displays information for the state occupation probabilities, state entry and exit distributions, and transition probabilities. Estimates of state occupation probability are displayed with corresponding variance estimates and confidence intervals (denoted `lower.ci` and `upper.ci` in the output). The default settings give these estimates to three decimal places for key percentile event times (minimum, maximum, 25th percentile, median, and 75th percentile). The `summary` method also displays estimates of the normalized state entry and exit time distributions (`entry.norm` and `exit.norm`, respectively), as well as the corresponding subdistributions (`entry.sub` and `exit.sub`).

The `summary` method also provides summary information for each allowed transition in the system. Estimates of the transition probabilities are given along with estimates of the variance and corresponding confidence intervals (`lower.ci` and `upper.ci`). In addition, the risk sets (`n.risk`), number of transitions (`n.event`, for transitions from one state to a different state), and number still remaining (`n.remain`, for transitions into the same state) are all displayed based on the counting process calculations in Andersen *et al.* (1995). Additionally (using `DS = TRUE`), users can display the weighted counting processes based on the formulas in Datta and Satten (2001) (`n.risk.K`, `n.event.K`, and `n.remain.K`, respectively).

```
R> summary(ex2, digits = 2)
```

State Occupation Information:

State 1

time	estimate	variance	lower.ci	upper.ci
0.07	1.00	9.8e-06	0.99	1.00
0.46	0.66	5.6e-04	0.62	0.71
0.71	0.39	6.0e-04	0.34	0.44
0.96	0.16	3.6e-04	0.12	0.19
2.38	0.00	0.0e+00	0.00	0.00

State 2

time	estimate	variance	lower.ci	upper.ci
0.07	0.0031	9.8e-06	0.00	0.0093
0.46	0.1363	2.9e-04	0.10	0.1697
0.71	0.1993	3.8e-04	0.16	0.2378
0.96	0.1649	3.3e-04	0.13	0.2006
2.38	0.0000	0.0e+00	0.00	0.0000

State 3

time	estimate	variance	lower.ci	upper.ci
0.07	0.00	0.0e+00	0.00	0.00
0.46	0.20	4.0e-04	0.16	0.24
0.71	0.41	5.9e-04	0.36	0.46
0.96	0.68	5.5e-04	0.63	0.72
2.38	1.00	-1.7e-18	NaN	NaN

State Entry Distribution Information:

State 3

time	entry.sub	entry.norm
0.07	0.00	0.00
0.46	0.20	0.20
0.71	0.41	0.41
0.96	0.68	0.68
2.38	1.00	1.00

State Exit Distribution Information:

State 1

time	exit.sub	exit.norm
0.07	0.0031	0.0031
0.46	0.3375	0.3375
0.71	0.6088	0.6088
0.96	0.8424	0.8424
2.38	1.0000	1.0000

Transition Probability Information:

Transition 1 -> 1

time	estimate	variance	lower.ci	upper.ci	n.risk	n.remain
0.07	1.00	9.8e-06	0.99	1.00	319	318
0.46	0.66	5.6e-04	0.62	0.71	269	268
0.71	0.39	6.0e-04	0.34	0.44	152	151
0.96	0.16	3.6e-04	0.12	0.19	53	53
2.38	0.00	0.0e+00	0.00	0.00	0	0

Transition 1 -> 2

time	estimate	variance	lower.ci	upper.ci	n.risk	n.event
0.07	0.0031	9.8e-06	0.00	0.0093	319	1
0.46	0.1363	2.9e-04	0.10	0.1697	269	1
0.71	0.1993	3.8e-04	0.16	0.2378	152	0
0.96	0.1649	3.3e-04	0.13	0.2006	53	0
2.38	0.0000	0.0e+00	0.00	0.0000	0	0

Transition 1 -> 3

time	estimate	variance	lower.ci	upper.ci	n.risk	n.event
0.07	0.00	0.0e+00	0.00	0.00	319	0
0.46	0.20	4.0e-04	0.16	0.24	269	0
0.71	0.41	5.9e-04	0.36	0.46	152	1
0.96	0.68	5.5e-04	0.63	0.72	53	0
2.38	1.00	-1.7e-18	NaN	NaN	0	0

Transition 2 -> 2

time	estimate	variance	lower.ci	upper.ci	n.risk	n.remain
0.07	1.00	0.0000	1.00	1.00	0	0
0.46	0.60	0.0142	0.37	0.84	60	60
0.71	0.38	0.0065	0.22	0.54	88	88
0.96	0.19	0.0020	0.11	0.28	74	73
2.38	0.00	0.0000	0.00	0.00	1	0

Transition 2 -> 3

time	estimate	variance	lower.ci	upper.ci	n.risk	n.event
0.07	0.00	0.0e+00	0.00	0.00	0	0
0.46	0.40	1.4e-02	0.16	0.63	60	0
0.71	0.62	6.5e-03	0.46	0.78	88	0
0.96	0.81	2.0e-03	0.72	0.89	74	1
2.38	1.00	-8.2e-19	NaN	NaN	1	1

Confidence intervals provided by default are 95% linear confidence intervals, but the user may change the confidence level to either 90% or 99% by changing the `ci.level` argument or apply a transformation of "log", "cloglog" or "log-log" by changing the `ci.trans` argument. The user may change the number of significant digits through the `digits` argument.

Summary information for all event times in the data set are displayed using the `all = TRUE` argument, while summary information at specific timepoints can be requested using the `times` argument.

```
R> summary(ex2, all = TRUE)
```

Users can customize the display of information by using the logical arguments `stateocc`, `trans.pr`, and `dist` to select whether the state occupation probabilities, transition probabilities, and state entry and exit time distributions are displayed. For example, to display only the state occupation probabilities use the following code:

```
R> summary(ex2, trans.pr = FALSE, dist = FALSE)
```

3.3. An example of state dependent censoring

State dependent censoring in `msSurv` will be illustrated using data on 136 cancer patients who received bone marrow transplants found in [Klein and Moeschberger \(2010\)](#). Following

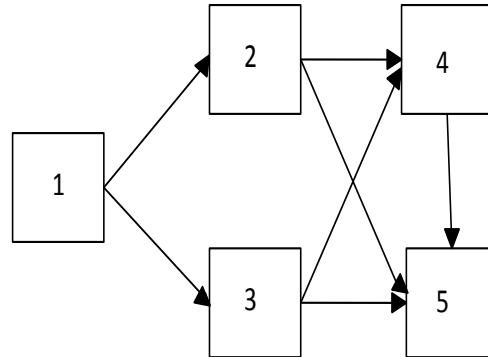


Figure 2: The five state model for cancer patients who received bone marrow transplants, discussed in Section 3.3.

Datta and Satten (2002), we defined five states based on platelets returning to a normal level, presence of acute graft-versus-host disease (GVHD), and onset of chronic GVHD.

State 2 is entered when acute GVHD develops before the patients platelet level returns to normal. State 3 is entered if the patient's platelet level returns to normal before acute GVHD develops. State 4 is for patients who have both normal platelet levels and acute GVHD while state 5 is for those patients who have chronic GVHD. Patients transition to either state 2 or state 3 and then either state 4 or state 5, as depicted in Figure 2. Patients do not necessarily progress to state 5 as they may remain in any state for any amount of time. Those patients who died or experienced relapse were considered censored for this example.

Our analysis allows for the censoring hazards to vary by state, since relapse or death can be predicted by the different (immunologic) states defined in this system. The cumulative censoring hazard for each state is estimated as the Nelson-Aalen estimator for censoring at each state and used to weigh the counting processes.

We open the data set `bmt` from the `KMsurv` (Yan 2010) package and form a data frame with column names `"id"`, `"stop"`, `"start.stage"`, and `"end.stage"`. (Code is suppressed here but is available in the accompanying R file.)

The first few lines of data are

```
R> if(exists("data.sdc")) head(data.sdc)
```

	id	stop	start.stage	end.stage	
	21	1	13	1	3
	199	1	67	3	4
	273	1	121	4	5
	22	2	18	1	3


```

218 2 139          3          5
23  3  12          1          3

```

Now we specify the tree structure.

```

R> Nodes <- c("1", "2", "3", "4", "5")
R> Edges <- list("1" = list(edges = c("2", "3")),
+              "2" = list(edges = c("4", "5")),
+              "3" = list(edges = c("4", "5")),
+              "4" = list(edges = c("5")),
+              "5" = list(edges = NULL))
R> deptree <- new("graphNEL", nodes = Nodes, edgeL = Edges,
+               edgemode = "directed")

```

Next we will call the `msSurv` function to perform nonparametric estimation on the data. Since the data is subject to state dependent censoring, we must add the argument `cens.type = "dep"` to the call of `msSurv`. Bootstrapping is used to estimate the variance for transition probabilities for state dependent censored data and may be specified using the `bs` argument. The default number of bootstraps is 200 and that is used for this example.

```

R> if(exists("data.sdc"))
+   DepEx <- msSurv(data.sdc, deptree, cens.type = "dep", bs = TRUE)

```

Entry distributions calculated for states 5 .

Exit distributions calculated for states 1 .

Using bootstrap to calculate variances ...

The results are displayed graphically using the `plot` method. The `plot` method takes `msSurv` objects and produces plots of estimated quantities using functions available in the `lattice` package (Sarkar 2008). The plots of the state occupation probabilities for every state in the system are produced by default with their corresponding 95% linear confidence intervals. For state dependent censoring, these confidence intervals are based on variances obtained through bootstrapping. Each state is plotted separately in a single panel. The results are given in Figure 3.

```

R> if(exists("DepEx")) plot(DepEx)

```

Users may also specify specific states to be plotted using the `states` argument. For example, to plot the state occupation probabilities for state 2, the user would type

```

R> if(exists("DepEx")) plot(DepEx, state = "2")

```

To plot the state occupation probabilities for states 2 and 3, use

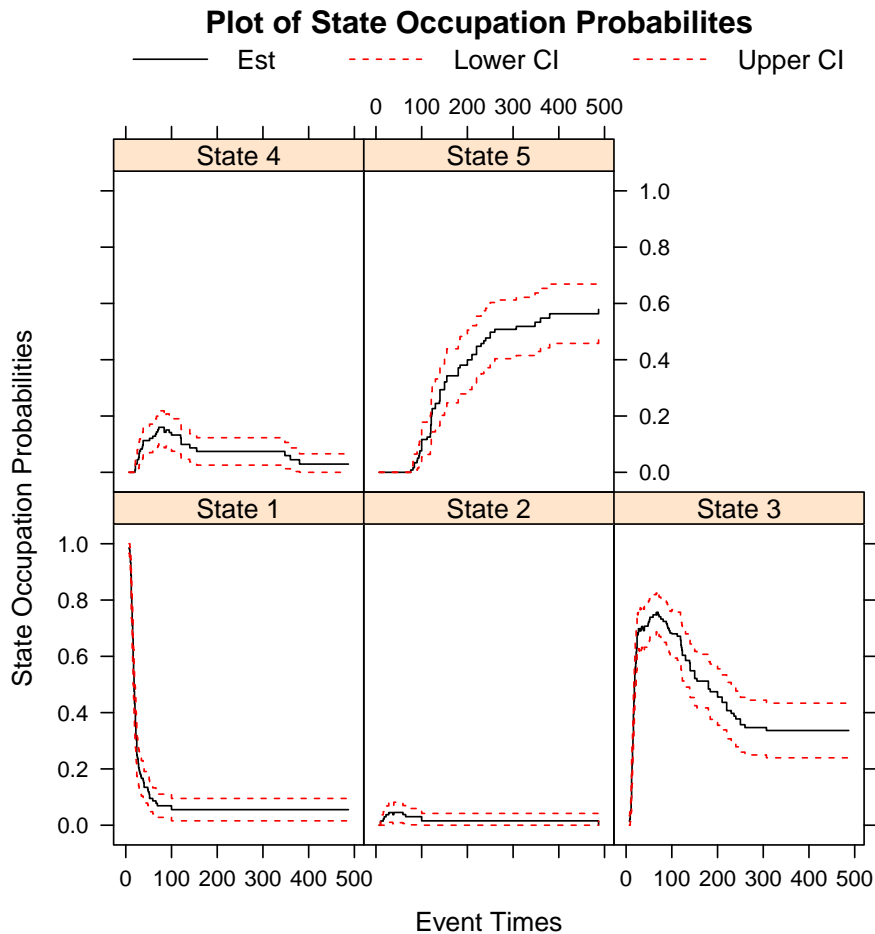


Figure 3: Plot of state occupation probability estimates and corresponding confidence intervals for each state in the system illustrated in Figure 2.

```
R> if(exists("DepEx")) plot(DepEx, state = c("2", "3"))
```

Confidence intervals may be altered using the `ci.level` and `ci.trans` arguments which allow the user to specify a different confidence level (0.9 or 0.99) or a different transformation ("log", "log-log", and "cloglog", as described in Section 2.5). Confidence intervals are omitted from the plots using the `CI = FALSE` argument.

The `plot.type` argument is used to change the plotted estimators. As previously mentioned, the default is "stateocc" for the state occupation probabilities, but the user may create plots for the transition probabilities ("transprob"), normalized state entry and exit distributions ("entry.norm" and "exit.norm", respectively), and state entry and exit subdistribution functions ("entry.sub" and "exit.sub", respectively). The state entry time distributions are plotted for all states except the initial state by default, but users may specify specific states using the "states" argument. The state exit time distributions are plotted for all non-terminal states by default, but specific non-terminal states may be requested by the user. We included the argument `bs = TRUE` in the call of `msSurv`, so that confidence interval estimates

of the state entry and exit distributions will be plotted.

The default plot produced for transition probabilities plots estimates and confidence intervals for all possible transitions in the system.

```
R> if(exists("DepEx"))
+   plot(DepEx, plot.type = "transprob")
```

The resulting plot is given in Figure 4.

Users may specify specific transitions using the `trans` argument. For example, if the user wants to plot the transition probabilities for the transitions out of state 1, say the "1 2" and "1 3" transitions, they would type

```
R> if(exists("DepEx"))
+   plot(DepEx, plot.type = "transprob", trans = c("1 2", "1 3"))
```

For state entry / exit distributions, in the current example the conditions given for ensuring proper calculation of these distributions (see Section 2.4) are only satisfied for states 1 and 5. To calculate state entry / exit distributions for the transient states, say states 2 and 3, we need to expand the model to differentiate between paths to states 4 and 5 that go through state 2 and paths that go through state 3. The expanded system is plotted in Figure 5, where states 4 and 5 have been relabeled as 4-2 and 5-2 for paths that go through state 2 and 4-3 and 5-3 for paths that go through state 3. The data needs to be modified correspondingly to incorporate these additional states. Note that individuals who start in state 4 are dropped from this model, as these individuals have no impact on the entry / exit distributions from states 2 and 3. The updated tree and data are in objects `deptree2` and `data.sdc2`, respectively.

The updated model is fitted below, and the normalized state entry distribution plots for states 2 and 3 over the time range of 0 to 100 days are given in Figure 6. The corresponding normalized exit distributions are plotted in Figure 7. By default (i.e., with the `states="ALL"` argument) all states where calculation of entry / exit distributions is valid are plotted.

```
R> if(exists("data.sdc2"))
+   DepEx2 <- msSurv(data.sdc2, deptree2, cens.type = "dep", bs = TRUE)
```

Entry distributions calculated for states 2 5-2 3 5-3 .

Exit distributions calculated for states 1 2 3 .

Using bootstrap to calculate variances ...

```
R> ## Now plot normalized entry distributions for states 2, 3
R> if(exists("DepEx2"))
+   plot(DepEx2, plot.type = "entry.norm", states = c("2", "3"),
+        xlim = c(0, 100))
```

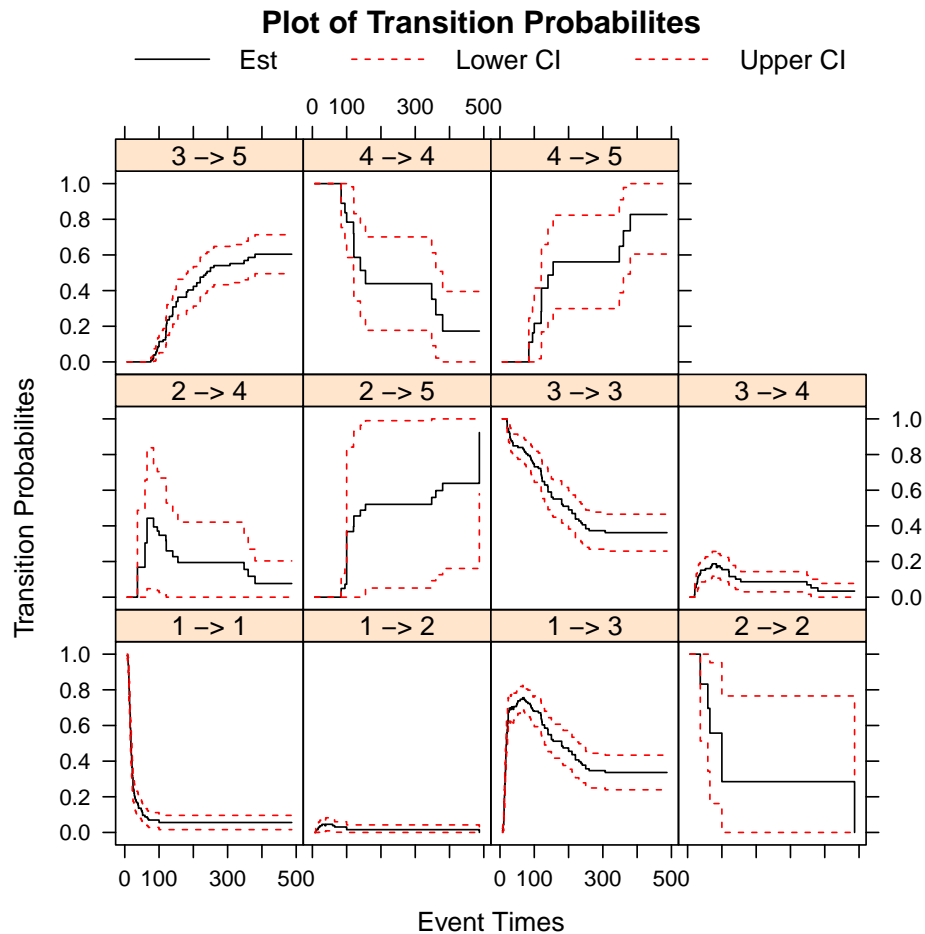


Figure 4: Plot of transition probability estimates and their corresponding confidence intervals for the system illustrated in Figure 2.

```
R> ## Now plot normalized exit distributions for states 2, 3
R> if(exists("DepEx2"))
+   plot(DepEx2, plot.type = "exit.norm", states = c("2", "3"))
R> ## Note that while normalized entry distributions will by
R> ## definition -> 1 as t -> infity, the normalized exit distributions
R> ## need not do this (e.g., if some individuals are never observed to
R> ## transition out of that state)
```

To illustrate how to integrate the information in the various estimates, we describe a clinical interpretation of the data. Starting from state 1, nearly all of the patients transition to either state 2 (acute GVHD) or state 3 (normal platelets) by day 100 (c.f. Figures 3 and 6). By 70 days post-transplant, roughly 75% of the patients have had their platelet levels return to normal before development of acute GVHD (Figure 3). The normalized entry time distribution for state 3 reaches approximately 1.0 by 100 days, indicating patients whose platelet levels return to normal prior to development of acute GVHD tend to do so within the first 100

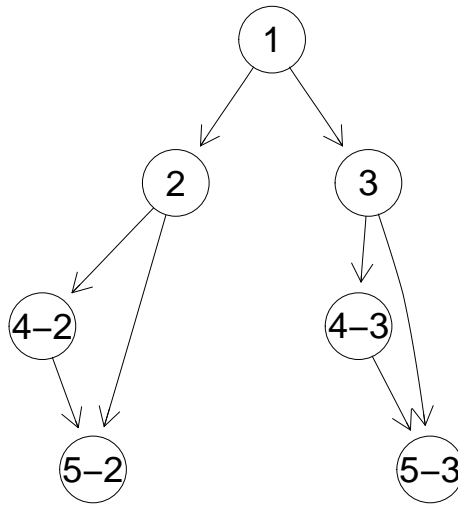


Figure 5: Expanded model discussed in Section 3.3, which differentiates between paths to states 4 and 5 that go through state 2 ($2 \rightarrow 4-2$ and $2 \rightarrow 5-2$) and state 3 ($3 \rightarrow 4-3$ and $2 \rightarrow 5-3$).

days. About 40% of these patients never develop GVHD (remain in state 3, Figure 7), with a small percentage subsequently developing acute GHVD ($3 \rightarrow 4$ transition in Figure 4) and 60% eventually going on to develop chronic GVHD ($3 \rightarrow 5$ transition in Figure 4). A much smaller percentage of patients first develop acute GVHD ($1 \rightarrow 2$ transition in Figure 4), with about 40% of these having their platelet levels returning to normal by 70-80 days (peak of $2 \rightarrow 4$ transition in Figure 4). All of the patients who develop acute GVHD first eventually go on to develop chronic GVHD ($2 \rightarrow 5$ transition in Figure 4).

3.4. An illustration of state-dependent censoring using simulated data

In the previous example based on real data the dependency of the censoring distribution on the state in the system was not too heavy, and the difference between estimates assuming state-dependent and state-independent censoring was relatively minor (data not shown). In this example we illustrate the magnitude of the difference between the two estimates that can occur using simulated data in the same spirit as [Datta and Satten \(2002\)](#). Data were simulated from a three state irreversible illness-death model, where all individuals begin in state 1 (well) at time 0. Individuals in state 1 may subsequently transition to state 2 (illness) or state 3 (death). Individuals with illness remain in state 2 until death (entry into state 3). For each individual, a vector $W = (W_1, W_2, W_3)$ was generated from a trivariate normal distribution with mean vector $(1.81, 1.7, 1.7)$ and variance-covariance matrix $0.393 \times (0.01\mathbf{I} + 0.91\mathbf{1}\mathbf{1}^T)$, where \mathbf{I} is the identity matrix and $\mathbf{1}$ is a 3×1 vector of ones. The values of W are (log) latent transition times with the following possible transition paths:

$W_1 < W_2$: Death occurs before illness at time $\exp(W_1)$

$W_2 < W_1$: Illness occurs at time $\exp(W_2)$ and death at time $\exp(W_2) + \exp(W_3)$

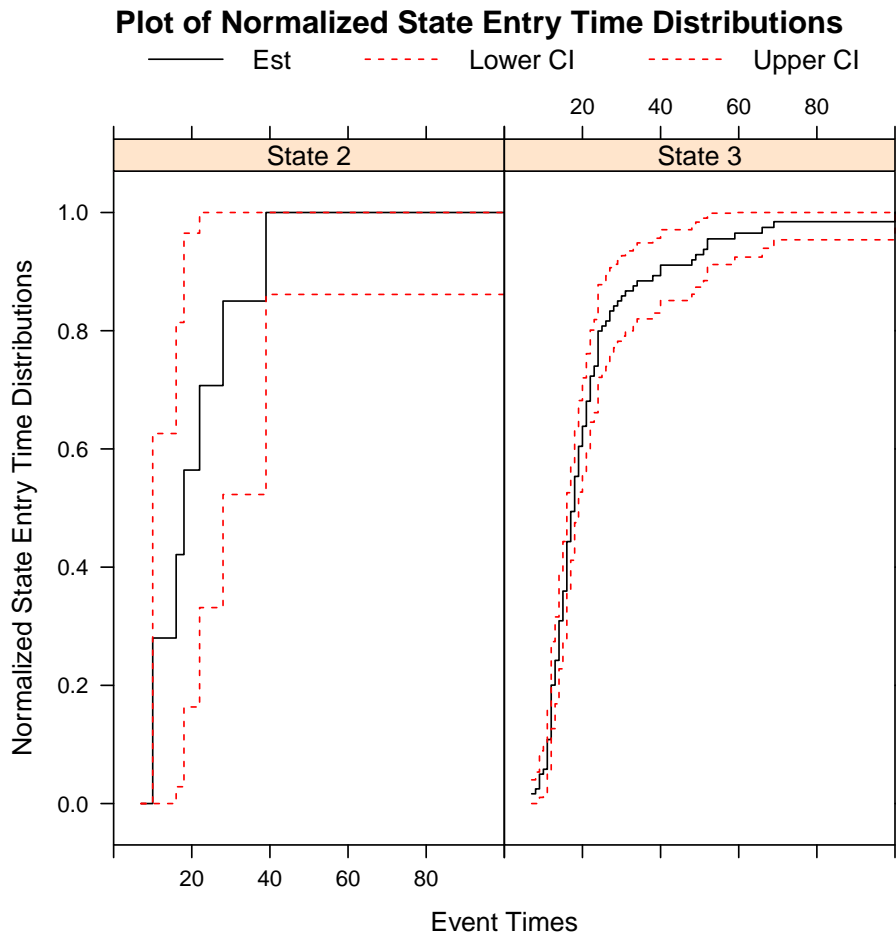


Figure 6: Plot of normalized state entry time distribution estimates and corresponding confidence intervals for states 2 and 3 in the system initially illustrated in Figure 2 and expanded in Figure 5.

The hazard for censoring was state dependent with censoring times following a Weibull distribution. In state 1, the censoring hazard was $t^{1/4}/60$, while in state 2 the censoring hazard was $t/50$. For each observation, a state 1 censoring time C_1 was generated and if $C_1 < \min\{\exp(W_1), \exp(W_2)\}$ the observation was censored in state 1. If $\exp(W_2) < C_1$ and if state 2 was entered, then a state 2 censoring time was generated from the Weibull distribution conditional on $C_2 > \exp(W_2)$. Simulations were based on a single data set of 1000 subjects. Figure 8 illustrates the state 2 occupation probability estimates based on the complete (non-censored) data, the Aalen-Johansen estimates, and the Datta-Satten estimates assuming state-dependent censoring. The Datta-Satten estimates closely follow the empirical estimates based on non-censored data, while the Aalen-Johansen estimates deviate appreciably.

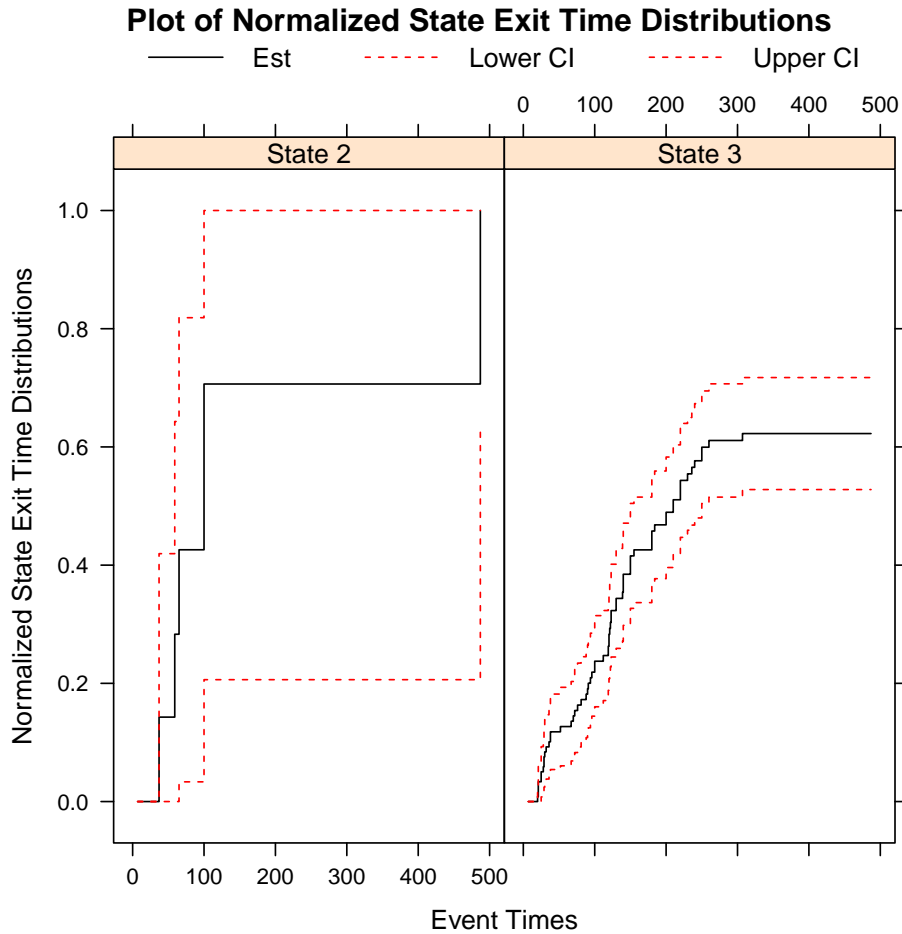


Figure 7: Plot of normalized state exit time distribution estimates and corresponding confidence intervals for states 2 and 3 in the system initially illustrated in Figure 2 and expanded in Figure 5.

4. Discussion

We present a comprehensive R package for nonparametric estimation of general multistate models subject to independent right censoring and possibly left truncation. This package computes the marginal state occupation probabilities and state entry and exit time distributions for possibly non-Markov models. For a Markov model, the R package, **msSurv** (Ferguson *et al.* 2012), also calculates and returns the marginal integrated transition hazard and the hazard rate functions, as well as the transition probability matrix between any two states. Motivated by Datta and Satten (2001), **msSurv** also performs nonparametric estimation for state dependent right censored and possibly left truncated data. Currently no other packages available on CRAN calculate the state entry and exit time distributions for censored multistate data or calculate nonparametric estimates for data subject to state dependent censoring. Pointwise confidence intervals for state occupation probability and transition probability matrices are obtained and returned for independent censoring from closed-form variance estimators and for

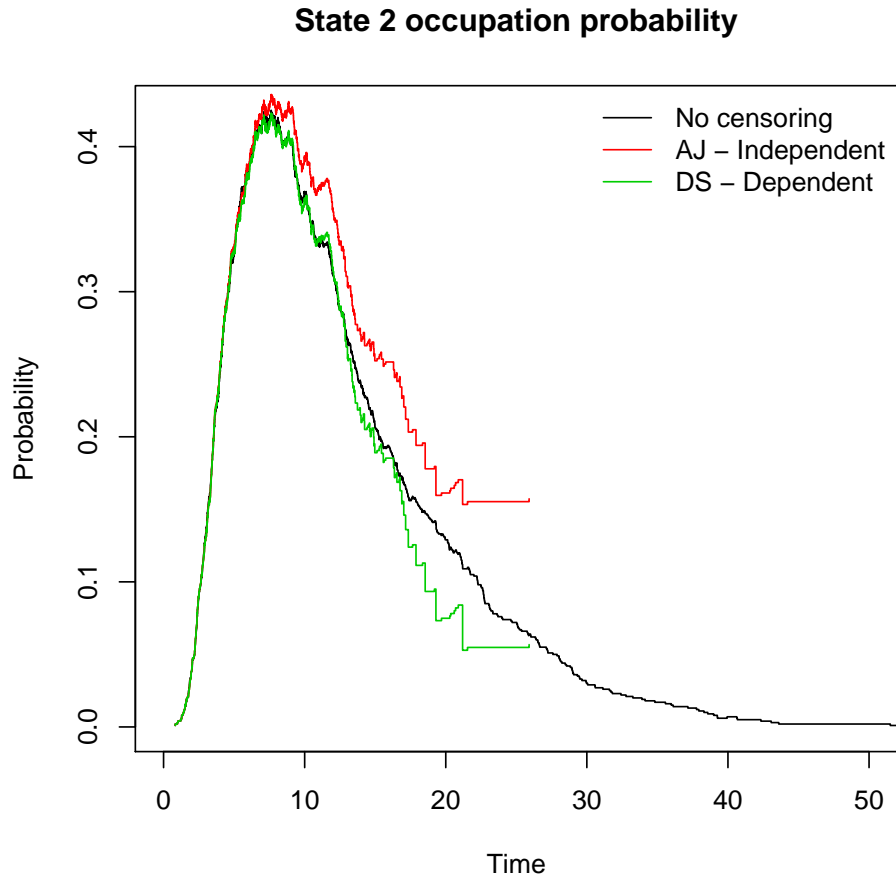


Figure 8: Plot of state 2 occupation probabilities for the three-state irreversible illness-death model in Section 3.4. The Datta-Satten (DS) estimator closely follows the empirical estimator based on complete (non-censored) data, while the Aalen-Johansen (AJ) estimator deviates noticeably.

state dependent censoring using the bootstrap. Pointwise confidence intervals for state entry and exit time distributions are obtained using the bootstrap. The bootstrap is also used to find variance estimators for state occupation probabilities of data subject to state dependent censoring. The **msSurv** package provides functions to find the state occupation probabilities at a specific time t and to find the transition probabilities between any two times s and t ($s \leq t$). Package **msSurv** is written using S4 classes and methods. Methods are available to print, plot, and summarize the **msSurv** objects.

The **msSurv** package has a number of limitations that will be improved upon in future releases. Currently state dependent censoring is the only censoring scheme incorporated in **msSurv**, future expansion could include incorporating general covariates into the (dependent) censoring mechanism. The package could also be extended to include estimation for current status data and eventually interval censored data (Datta and Sundaram 2006; Lan and Datta 2010). Other

extensions include calculations of the state waiting time distributions (Satten and Datta 2002) and allowing estimation for recurrent event models.

References

- Aalen OO (1978). “Nonparametric Inference for a Family of Counting Processes.” *The Annals of Statistics*, **6**(4), 701–726.
- Aalen OO (1980). “A Model for Non-Parametric Regression Analysis of Counting Processes.” In W Klonecki, A Kozek, J Rosiskipp (eds.), *Lecture Notes in Statistics-2: Mathematical Statistics and Probability Theory*, pp. 1–25. Springer-Verlag, New York.
- Aalen OO, Johansen S (1978). “An Empirical Transition Matrix for Nonhomogeneous Markov Chains Based on Censored Observations.” *Scandinavian Journal of Statistics*, **5**(3), 141–150.
- 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.
- Andersen PK, Borgan Ø, Gill RD, Keiding N (1995). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.
- Bie O, Borgan O, Liestol K (1987). “Confidence Intervals and Confidence Bands for the Cumulative Hazard Rate Function and their Small Sample Properties.” *Scandinavian Journal of Statistics*, **14**(3), 221–233.
- Borgan O, Liestol K (1990). “A Note on Confidence Intervals and Bands for the Survival Curve Based on Transformations.” *Scandinavian Journal of Statistics*, **17**(1), 35–41.
- Datta S, Ferguson AN (2011). “Nonparametric Estimation of Marginal Temporal Functionals in a Multistate Model.” In A Lisnianski, I Frenkel (eds.), *Recent Advances in System Reliability: Signatures, Multi-State Systems and Statistical Inference*. Springer.
- Datta S, Satten GA (2000). “Estimating Future Stage Entry and Occupation Probabilities in a Multistage Model Based on Randomly Right-Censored Data.” *Statistics and Probability Letters*, **50**(1), 89–95.
- 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 and Probability Letters*, **55**(4), 403 – 411.
- Datta S, Satten GA (2002). “Estimation of Integrated Transition Hazards and Stage Occupation Probabilities for Non-Markov Systems Under Dependent Censoring.” *Biometrics*, **58**(4), 792–802.
- Datta S, Sundaram R (2006). “Nonparametric Estimation of State Occupation Probabilities in a Multistate Model with Current Status Data.” *Biometrics*, **62**, 829–837.

- de Wreede LC, Fiocco M, Putter H (2011). “**mstate**: An R Package for the Analysis of Competing Risks and Multistate Models.” *Journal of Statistical Software*, **38**(7), 1–30.
- 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. URL <http://www.jstatsoft.org/v50/i14/>.
- Gentleman R, Whalen E, Huber W, Falcon S (2010). **graph**: A Package To Handle Graph Data Structures. Bioconductor package version 1.34.0, URL <http://www.bioconductor.org/packages/release/bioc/html/graph.html>.
- Jackson C (2011). “Multi-State Models for Panel Data: The **msm** Package for R.” *Journal of Statistical Software*, **38**(8), 1–28.
- Kalbfleisch JD, Prentice RL (2002). *The Statistical Analysis of Failure Time Data*. 2nd edition. John Wiley and Sons, New York.
- Klein JP, Moeschberger ML (2010). *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd edition. Springer, New York.
- Lan L, Datta S (2010). “Nonparametric Estimation of State Occupation, Entry and Exit Times with Multistate Current Status Data.” *Statistical Methods in Medical Research*, **19**, 147–165.
- Meira-Machado L, Roca-Pardinas J (2011). “**p3state.msm**: Analyzing Survival Data from an Illness-Death Model.” *Journal of Statistical Software*, **38**(3), 1–18.
- Nelson W (1972). “Theory and Applications of Hazard Plotting for Censored Failure Data.” *Technometrics*, **14**(4), 945–965.
- Pepe MS (1991). “Inference for Events with Dependent Risks in Multiple End Point Studies.” *Journal of American Statistical Association*, **86**(415), 770–778.
- R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Robins J, Rotnitzky A (1992). “Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers.” In N Jewell, K Dietz, V Farewell (eds.), *AIDS Epidemiology - Methodological Issues*, pp. 297–331. Birkhauser, Boston.
- Sarkar D (2008). **Lattice**: *Multivariate Data Visualization with R*. Springer, New York. ISBN 978-0-387-75968-5, URL <http://lmdvr.r-forge.r-project.org>.
- Satten GA, Datta S (2002). “Marginal Estimation for Multistage Models: Waiting Time Distributions and Competing Risk Analyses.” *Statistics in Medicine*, **21**(1), 3–19.
- Schulgen G, Schumacher M (1996). “Estimation of Prolongation of Hospital Stay Attributable to Nosocomial Infections: New Approaches Based on Multistate Models.” *Lifetime Data Analysis*, **2**(3), 219–40.
- Thomas DR, Grunkemeier GL (1975). “Confidence Interval Estimation of Survival Probabilities for Censored Data.” *Journal of American Statistical Association*, **70**(352), 865–871.

Wrangler M, Beyersmann J, Schumacher M (2006). “**changeLOS**: An R-Package for Change in Length of Hospital Stay Based on the Aalen-Johansen Estimator.” *R News*, **6**(2), 31–35.

Yan J (2010). *KMsurv: Data sets from Klein and Moeschberger (1997), Survival Analysis*. R package version 0.1-4, URL <http://CRAN.R-project.org/package=KMsurv>.

Affiliation:

Somnath Datta
Department of Bioinformatics and Biostatistics
School of Public Health and Information Sciences
University of Louisville
485 E. Gray St.
Louisville, KY, USA 40292
E-mail: somnath.datta@louisville.edu