SmoothHazard

An R package for fitting regression models to interval-censored observations of illness-death models

Touraine, Célia; Gerds, Thomas A.; Joly, Pierre

Published in:
Journal of Statistical Software

DOI:
10.18637/jss.v079.i07

Publication date:
2017

Document version
Publisher's PDF, also known as Version of record

Journal of Statistical Software

July 2017, Volume 79, Issue 7. *doi: 10.18637/jss.v079.i07*

SmoothHazard: An R Package for Fitting Regression Models to Interval-Censored Observations of Illness-Death Models

Célia Touraine
University of Bordeaux

Thomas A. Gerds
University of Copenhagen

Pierre Joly
University of Bordeaux

Abstract

The irreversible illness-death model describes the pathway from an initial state to an absorbing state either directly or through an intermediate state. This model is frequently used in medical applications where the intermediate state represents illness and the absorbing state represents death. In many studies, disease onset times are not known exactly. This happens for example if the disease status of a patient can only be assessed at follow-up visits. In this situation the disease onset times are interval-censored. This article presents the **SmoothHazard** package for R. It implements algorithms for simultaneously fitting regression models to the three transition intensities of an illness-death model where the transition times to the intermediate state may be interval-censored and all the event times can be right-censored. The package parses the individual data structure of the subjects in a data set to find the individual contributions to the likelihood. The three baseline transition intensity functions are modelled by Weibull distributions or alternatively by M-splines in a semi-parametric approach. For a given set of covariates, the estimated transition intensities can be combined into predictions of cumulative event probabilities and life expectancies.

Keywords: illness-death model, interval-censored data, left-truncated data, survival model, smooth transition intensities, Weibull, penalized likelihood, M-splines.

1. Introduction

The irreversible illness-death model is a multi-state model which has many applications in various areas of research, for example in the medical field. The model describes the transitions from an initial state (e.g., alive and disease-free) to an absorbing state (e.g., death) either directly or via an intermediate state (e.g., disease, Figure 1). The transition intensities α_{01}, α_{02}, and α_{12} are positive functions of time which can also depend on covariates.

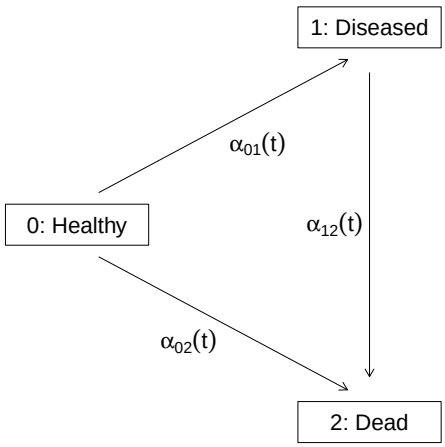

Figure 1: The irreversible illness-death model.

In some applications it happens for some or all subjects that the transition times from the initial state to the intermediate state are interval-censored. This occurs for example when the status of the intermediate state can only be determined at a sequence of visit times. In this case, if a subject is diagnosed as diseased at one of the visit times, say R, then it is only known that the subject was last seen disease-free at the previous visit time, say L, and hence the time of the onset of the disease is interval-censored between L and R for this subject. Furthermore, both the process of visit times and the observation of the time of the transition into the absorbing state are usually right-censored, i.e., limited to the individual follow-up period of the subjects. This yields a rather complex general observational pattern, because for a subject who died without being diagnosed as diseased at earlier visit times, it may or it may not be possible to determine retrospectively if and when the subject became diseased between the last visit time and the time of death.

The **SmoothHazard** package (Touraine, Joly, and Gerds 2017) provides estimates of the baseline transition intensities and of covariate effects when the data fall into one of the 6 cases that are displayed in Figure 2. Thus, the case of left-truncated event times (delayed entry) is covered, as well as the case where for some subjects the transition time into the intermediate state is observed exactly and for others it is interval-censored. Finally, the special case is covered where for some or all subjects no intermediate information is available about the disease status such that it is only known whether or not the subjects became diseased between the start and the end of follow-up. The latter occurs in Figure 2 when $E = L$ and $R = \min(T, C)$ in cases 3 or 6.

The aim is to estimate covariate effects on the three transition intensities. To achieve this regression models are implemented which assume proportional transition intensities and a non-homogeneous Markov process. The user chooses between a fully parametric model where each of the baseline intensities is described by the parameters of a Weibull distribution and a semi-parametric model where the baseline intensities are left unspecified and approximated by M-splines. For the parametric model, the regression coefficients and Weibull parameters are estimated by maximizing the likelihood; for the semi-parametric model, the coefficients of the M-splines and the regression coefficients are estimated by maximizing a penalized likelihood. The **SmoothHazard** package then allows predictions of transition probabilities, cumulative

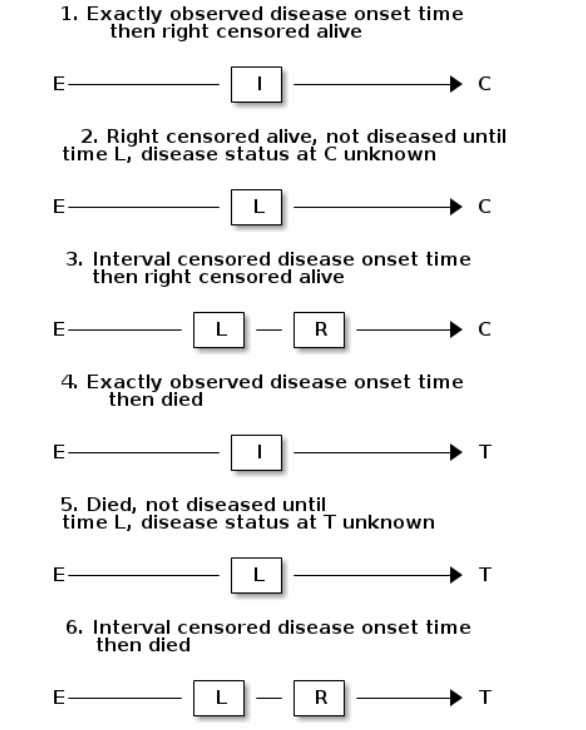

Figure 2: Observational patterns that are recognized by **SmoothHazard**. The letters I and T denote the transition times into the intermediate and absorbing state, respectively. The letters E and C denote the start and end of follow-up, respectively, and the letters L and R the visit times between which the transition into the intermediate happened.

probabilities of event and life expectancies to be obtained for a given set of covariates, based on estimated baseline transition intensities and on estimated covariate effects.

The **SmoothHazard** package is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=SmoothHazard`.

Comparison to other packages

If the exact transition times of an illness-death model are observed, standard procedures can be used to estimate transition intensities, regression coefficients and functionals thereof (Putter, Fiocco, and Geskus 2007). In particular, in R (R Core Team 2017) the packages **survival** (Therneau and Grambsch 2000; Therneau 2017) and **rms** (Harrell 2017) estimate the regression coefficients using the Cox partial likelihood (Cox 1975) without the need to model the baseline intensities. Besides, the package **simPH** (Gandrud 2015) is useful to simulate and plot quantities of interest and their uncertainty for coefficient estimates from Cox models (even interactive and nonlinear effects). Moreover, a number of packages facilitate the use of multi-state models and illness-death models. The packages **etm** (Allignol, Schumacher, and Beyersmann 2011) and **mstate** (de Wreede, Fiocco, and Putter 2010, 2011; Putter *et al.* 2007) can be used to estimate transition probabilities for the Aalen-Johansen estimator but the **etm** package does not account for the influence of covariates. The **p3state.msm** package (Meira-Machado and Roca-Pardiñas 2011) obtains non-Markov estimates for the transition

probabilities. The **TPmsm** (Araújo, Meira-Machado, and Roca-Pardiñas 2014) package implements several estimators for the transition probabilities, including the Aalen-Johansen estimator and estimators that are consistent even without Markov assumption or in case of dependent censoring.

However, when transition times to the intermediate event are interval-censored, it is then generally not possible to arrive at consistent estimates with the software provided by the packages listed above. Indeed, a common approach for handling subjects who died with unknown disease status consists of artificially ending their follow-up at the last time they were seen without disease and subsequently treating them as right-censored. However, this approach can lead to a systematic bias in the estimates of transition intensities and of regression coefficients (Joly, Commenges, Helmer, and Letenneur 2002; Leffondré, Touraine, Helmer, and Joly 2013). The bias will be especially pronounced if the risk of death is higher for diseased subjects than the risk of death for disease-free subjects.

To our knowledge, there is only one package that focus on fitting illness-death models to interval-censored data, the **coxinterval** package (Boruvka and Cook 2015) that implements a sieve estimator for the Cox regression model. The **msm** package (Jackson 2011) is able to fit Markov multi-state models to panel data where the status of the subjects is known at a finite series of inspection times. As a special case the setting includes the illness-death model and it can be used with interval-censored disease times and exact death times. However, in this package the likelihood is calculated using the Kolmogorov differential equations that relate the transition probabilities and the transition intensities and in order for there to be an analytic solution a time-homogeneity assumption is made where all transition intensities are constant or piecewise-constant between two successive observation times.

The **SmoothHazard** package fits also survival models to data that might be interval-censored. However, other packages can be used in this case. The available packages include the **ICsurv** package (McMahan and Wang 2014) that implements a semi-parametric proportional hazard regression model, the **intcox** package that implements an algorithm for extending the Cox regression model (Pan 1999), the **MIICD** package (Delord 2016) that implements a multiple imputation approach to Cox regression and the **coarseDataTools** package that implements the parametric accelerated failure time model (Reich, Lessler, Cummings, and Brookmeyer 2009; Reich, Lessler, and Azman 2016). The **interval** package (Fay and Shaw 2010) implements tests for comparing survival distributions for interval-censored data.

Outline of ***SmoothHazard***

To sum up, **SmoothHazard** fits survival models or illness-death models to data observed in either continuous times (leading to exactly observed or right-censored data) or in discrete times (leading to interval-censored or right-censored data). In the illness-death model, only the entry time into the intermediate state (diseased) can be interval-censored; the entry time into the absorbing state (death) is observed in continuous time. **SmoothHazard** produces smooth estimates for the hazard function (in a survival model) or for the transition intensity functions (in an illness-death model) by assuming Weibull or M-splines baseline function(s).

The main functions of **SmoothHazard** are

- `shr`: for fitting survival regression models based on possibly interval-censored times and right-censored times;

- `idm`: for fitting illness-death regression models based on possibly interval-censored disease times and right-censored times.

A fitted illness-death model as produced by `idm` can be used to calculate predictions with:

- S3 `predict` method for '`idm`' objects: for estimating transition probabilities, cumulative event probabilities and life expectancies for a given set of covariates.

In this paper, we focus on the main goal of the package which is fitting illness-death models to interval-censored data. Section 2 presents the model and the likelihood. Section 3 presents the estimation methods. Section 4 briefly presents predictions that can be made in an illness-death model. Section 5 provides some examples illustrating **SmoothHazard**. We discuss limitations and further extensions in Section 6.

2. Model and likelihood

We consider an illness-death process $X = (X(t), t \geq 0)$ which takes values in $\{0, 1, 2\}$ (Figure 1). Subjects are initially disease-free ($X(0) = 0$) and may become diseased (transition $0 \to 1$) and die (transition $1 \to 2$), or die directly without disease (transition $0 \to 2$). For more in-depth treatment of the topic, see Andersen and Keiding (2012) and Keiding (2014). X is assumed to be a non-homogeneous Markov process which means that the future evolution of the process $\{X(t), t > s\}$ depends on the current time s and only on the current state $X(s)$. Thus, the distribution of X is fully characterized by the set of transition probabilities:

$$p_{hl}(s,t) = \mathsf{P}(X(t) = l | X(s) = h), \qquad hl \in \{01, 02, 12\}.$$

The transition probabilities are related to the instantaneous transition intensities α_{hl} shown in Figure 1 by the relation:

$$\alpha_{hl}(t) = \lim_{\Delta t \to 0} \frac{p_{hl}(t, t + \Delta t)}{\Delta t}.$$

We introduce covariate effects separately for each transition through proportional transition intensities regression models which are a natural extension of the Cox proportional hazard model:

$$\alpha_{hl}(t|Z_{hli}) = \alpha_{0,hl}(t) \exp\{\beta_{hl}^\top Z_{hli}\}, \qquad hl \in \{01, 02, 12\}. \tag{1}$$

Here $\alpha_{0,hl}$ are baseline transition intensities, Z_{hli} are covariate vectors for subject i and β_{hl} are vectors of regression parameters for transition $h \to l$.

In the situation where the time to disease and the time to death are not interval-censored but either observed exactly or right-censored, the regression coefficients could be estimated by the partial likelihood method without the need to specify and estimate the baseline transition intensity functions $\alpha_{0,hl}(t)$. For interval-censored transition times to the intermediate state, the situation is more complex. It turns out that we have to estimate all parameters simultaneously and that we need a model for the baseline transition intensity functions. This can be seen by inspecting the likelihood function.

For subject i, denote the conditional disease-free survival function by

$$S(t|Z_{01i}, Z_{02i}) = e^{-A_{01}(t|Z_{01i}) - A_{02}(t|Z_{02i})},$$

where $A_{hl}(\cdot|Z_{hli})$ is the conditional cumulative intensity function of transition $h \to l$:

$$A_{hl}(t|Z_{hli}) = \int_0^t \alpha_{hl}(u|Z_{hli})du.$$

Note that the conditional probability of surviving time t given a transition into the intermediate state at time s is given by $\exp\{-A_{12}(t|Z_{12i}) + A_{12}(s|Z_{12i})\}$.

We allow that the event times are left-truncated, i.e., that subjects enter the study at the delayed entry time $E > 0$. The left-truncation condition $X(E_i) = 0$ implies that subject i has survived in state 0 until time E_i.

In addition to the covariate vectors $Z_{01i}, Z_{02i}, Z_{12i}$ we observe the vector $(E_i, L_i, R_i, \delta_{1i}, \tilde{T}_i, \delta_{2i})$ where $\tilde{T}_i = \min(T_i, C_i)$ is the minimum between the transition time into the absorbing state T_i and the right censoring time C_i and $\delta_{2i} = \mathbb{1}\{T_i \leq C_i\}$. Also, $\delta_{1i} = 1$ if we know for sure that subject i was diseased between E_i and $\tilde{T}_i$ and $\delta_{1i} = 0$ otherwise. The visit times L_i and R_i are defined by $E_i \leq L_i \leq R_i \leq \tilde{T}_i$ if $\delta_{1i} = 1$ and by $E_i \leq L_i \leq \tilde{T}_i, R_i = \infty$ if $\delta_{1i} = 0$. When the transition time into the intermediate state is observed exactly, we have $\delta_{1i} = 1$ and $L_i = R_i$. In the latter case we also denote I_i for the transition time into the intermediate state.

We now detail the likelihood contributions according to the different observational patterns shown in Figure 2 in the special case where there is no left-truncation i.e., $E_i = 0$. Left-truncated event times are taken into account by simply dividing the above likelihood contributions by the term $S(E_i|Z_{01i}, Z_{02i})$.

$$\begin{aligned}
\text{case 1:} \quad & \mathcal{L}_i = S(I_i|Z_{01i}, Z_{02i})\alpha_{01}(I_i|Z_{01i})\frac{e^{-A_{12}(C_i|Z_{12i})}}{e^{-A_{12}(I_i|Z_{12i})}} \\
\text{case 2:} \quad & \mathcal{L}_i = S(C_i|Z_{01i}, Z_{02i}) + \int_{L_i}^{C_i} S(u|Z_{01i}, Z_{02i})\alpha_{01}(u|Z_{01i})\frac{e^{-A_{12}(C_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}}du \\
\text{case 3:} \quad & \mathcal{L}_i = \int_{L_i}^{R_i} S(u|Z_{01i}, Z_{02i})\alpha_{01}(u|Z_{01i})\frac{e^{-A_{12}(C_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}}du \\
\text{case 4:} \quad & \mathcal{L}_i = S(I_i|Z_{01i}, Z_{02i})\alpha_{01}(I_i|Z_{01i})\frac{e^{-A_{12}(T_i|Z_{12i})}}{e^{-A_{12}(I_i|Z_{12i})}}\alpha_{12}(T_i|Z_{12i}) \\
\text{case 5:} \quad & \mathcal{L}_i = S(T_i|Z_{01i}, Z_{02i})\alpha_{02}(T_i|Z_{02i}) \\
& \qquad + \int_{L_i}^{T_i} S(u|Z_{01i}, Z_{02i})\alpha_{01}(u|Z_{01i})\frac{e^{-A_{12}(T_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}}\alpha_{12}(T_i|Z_{12i})du \\
\text{case 6:} \quad & \mathcal{L}_i = \int_{L_i}^{R_i} S(u|Z_{01i}, Z_{02i})\alpha_{01}(u|Z_{01i})\frac{e^{-A_{12}(T_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}}\alpha_{12}(T_i|Z_{12i})du
\end{aligned} \tag{2}$$

3. Estimation

The `idm` function computes estimates for the three baseline transition intensities and for the regression parameters using the Levenberg-Marquardt's algorithm (Levenberg 1944; Marquardt 1963) to maximize the (penalized) likelihood. The algorithm is a combination of a Newton-Raphson algorithm and a gradient descent algorithm (also known as the steepest descent algorithm). It has the advantage of being more robust than the Newton-Raphson algorithm while preserving its fast convergence property.

3.1. Parametric estimation

A Weibull parametrization for the baseline transition intensities is assumed in `idm`'s default estimation method:

$$\alpha_{0,hl}(t) = a_{hl}\ b_{hl}^{a_{hl}}\ t^{a_{hl}-1}, \qquad hl \in \{01, 02, 12\},$$

where a_{hl} and $\frac{1}{b_{hl}}$ are shape and scale parameters. The Weibull parameter estimates $\hat{a}_{hl}$ and $\hat{b}_{hl}$ and the vectors of regression parameter estimates $\hat{\beta}_{hl}$ are obtained simultaneously by maximizing the likelihood which is the product over the subjects' contributions according to (2):

$$\mathcal{L}(\beta_{01}, \beta_{02}, \beta_{12}, a_{01}, a_{02}, a_{12}, b_{01}, b_{02}, b_{12}) = \prod_{i=1}^{n} \mathcal{L}_i(\beta_{01}, \beta_{02}, \beta_{12}, a_{01}, a_{02}, a_{12}, b_{01}, b_{02}, b_{12}). \quad (3)$$

Confidence intervals for the regression parameters are obtained using standard errors estimated by inverting the Hessian matrix of the log-likelihood, that is the matrix of the second partial derivatives of $\log \mathcal{L}$ given in (3). Pointwise confidence intervals for the baseline transition intensities are obtained using a simulation-based approach explained in Section 4.1.

3.2. Semi-parametric estimation

In situations where it is suspected that the Weibull distribution does not fit the data very well one can think of extending the model and leaving the baseline intensity functions completely unspecified, as in the Cox regression model. Unfortunately, in interval-censored data there is no direct analogue to the partial likelihood and the Breslow estimator of the Cox model in right-censored data. The function `idm` implements a semi-parametric model where the three baseline transition intensities are approximated by linear combinations of M-splines. In this section we explain the basic steps of this approach.

The penalized likelihood

To control the smoothness of the estimated intensity functions, we penalize the log-likelihood by a term which specifies the curvature of the intensity functions. It is given by the square of the second derivatives. The penalized log-likelihood (pl) is defined as:

$$pl = l - \kappa_{01} \int \alpha_{01}''^{\,2}(u|Z_{01})du - \kappa_{02} \int \alpha_{02}''^{\,2}(u|Z_{02})du - \kappa_{12} \int \alpha_{12}''^{\,2}(u|Z_{12})du, \quad (4)$$

where l is the log-likelihood and κ_{01}, κ_{02} and κ_{12} are three positive parameters which control the trade-off between the data fit and the smoothness of the functions. It is proposed that the penalization parameters are chosen by maximizing a cross-validated likelihood score. Here, leave-one-out is appealing as the result does not depend on the random seed as it would, e.g., for 10-fold cross-validation. However, since leave-one-out requires as many maximizations of the likelihood as there are subjects in the data set, this can be computationally very expensive. To avoid extremely long run times we have implemented the following algorithm:

Step 1. We ignore the covariates and use a grid search method to find the values for the parameters $(\kappa_{01}, \kappa_{02}, \kappa_{12})$ based on an approximation of the leave-one-out log-likelihood score. This approximate score requires the computation of the Hessian matrix and

one step of the Newton-Raphson algorithm. Thus it reduces the number of calculations considerably. The approximate leave-one-out log-likelihood score is similar to an Akaike information criterion (for more details, see Commenges, Joly, Gégout-Petit, and Liquet 2007). This approach was proposed by O'Sullivan (1988) for survival models and studied by Joly *et al.* (2002) in an illness-death model with interval-censored data.

Step 2. We use the results of Step 1, i.e., the optimized value of $(\kappa_{01}, \kappa_{02}, \kappa_{12})$ to maximize the penalized likelihood (see Equation 4) with covariates. The parameters being maximized are the regression coefficients and the coefficients of the linear combinations of the M-splines defined below.

M-splines

We use linear combinations of M-spline basis functions which are positive splines and variants of B-splines. A family of M-spline functions of order k, $M_1, \ldots, M_n$ is defined by a set of m knots where $n = m + k - 2$ (Ramsay 1988). We consider only cubic M-splines of order $k = 4$. Denote by $t_{01} = (t_{01,1}, \ldots, t_{01,m_{01}})$ a sequence of m_{01} knots used for approximating α_{01} and by $t_{02} = (t_{02,1}, \ldots, t_{02,m_{02}})$ and $t_{12} = (t_{12,1}, \ldots, t_{12,m_{12}})$ similar sequences of knots for approximating α_{02} and α_{12} respectively. We denote by $M_{hl} = \{M_{hl,1}, \ldots, M_{hl,n_{hl}}\}$ the families of n_{hl} cubic M-splines, with $n_{hl} = m_{hl} + 2$ and for $hl \in \{01, 02, 12\}$. The baseline transition intensity $\alpha_{0,hl}$ is approximated using the following linear combination:

$$\tilde{\alpha}_{0,hl}(t) = \sum_{i=1}^{n_{hl}} (a_{hl,i})^2 M_{hl,i}(t),$$

where $a_{hl,i}$ are unknown parameters. The n_{hl} M-splines are integrated in order to produce a family of monotone splines, these are called I-splines. Thus, with each M-spline $M_{hl,i}$ we associate an I-spline $I_{hl,i}$:

$$I_{hl,i}(t) = \int_{t_{hl,1}}^{t} M_{hl,i}(u)du.$$

For given values of the parameters $a_{hl,i}$, we can approximate the cumulative baseline transition intensities A_{hl} by a linear combination of I-splines:

$$\tilde{A}_{0,hl}(t) = \sum_{i=1}^{n_{hl}} (a_{hl,i})^2 I_{hl,i}(t).$$

Because M-splines are non-negative, the positivity constraint on $(a_{hl,i})^2$ ensures that $\tilde{A}_{0,hl}$ is monotone increasing.

Confidence intervals of the regression parameters are obtained using estimated standard errors which are obtained by inverting the Hessian matrix of the penalized log-likelihood.

Confidence intervals for the baseline transition intensities $\alpha_{0,hl}(t)$ are obtained using the Bayesian approach proposed in O'Sullivan (1988) for survival analysis where the standard errors are estimated by $M_{hl}(t)^\top H_{hl}^{-1} M_{hl}(t)$, where H_{hl} denotes the specific part for transition from h to l of the Hessian matrix of the penalized log-likelihood.

4. Predictions

Often in illness-death models the functions of interest are the transition intensities. However, other quantities (transition probabilities, cumulative probabilities and life expectancies) which can be expressed in terms of the transition intensities (Touraine, Helmer, and Joly 2016) may provide additional information and have a more natural interpretation.

For example, given a set of covariates $Z_{01,i}, Z_{02,i}, Z_{12,i}$ for a subject i who is diseased at time s, one could be interested in the probability that the subject is still alive at some time $t > s$, or the subject's life expectancy. Given a set of covariates $Z_{01,j}, Z_{02,j}, Z_{12,j}$ for a subject j who is disease-free at time s, one could be interested in lifetime risk of disease or in healthy life expectancy (expected remaining sojourn time in the healthy state). Since these quantities can be written in terms of the transition intensities, **SmoothHazard** provides estimates of them using estimates of the transition intensities. Confidence intervals are calculated using the simulation-based method described in the following.

4.1. Confidence regions

A simulation based approach (Mandel 2013) is used to calculate confidence intervals for the transition intensities $\alpha_{hl}(t)$ in the parametric approach and for the other quantities of interest in both parametric and semi-parametric approaches. To briefly outline how it works, we generically denote by θ the vector of all the parameters that characterize the likelihood and by $\hat{\theta}$ the maximum (penalized) likelihood estimator. θ contains the Weibull parameters in the parametric model, the spline parameters in the semi-parametric model and the regression parameters in both models.

We assume asymptotic normality for the estimator $\hat{\theta}$ and denote by $\hat{V}_{\hat{\theta}}$ the estimated covariance matrix of $\hat{\theta}$. We consider a multivariate normal distribution with the parameter estimates as expectation and $\hat{V}_{\hat{\theta}}$ as covariance matrix. We generate n vectors ($n = 2000$ in practice) from this distribution: $\theta^{(1)}, \ldots, \theta^{(n)}$. Based on them, we can calculate n values for the transition intensities: $\alpha_{hl}^{(1)}(t)$, $\ldots$, $\alpha_{hl}^{(n)}(t)$, and therefore n values for any quantity of interest written in terms of the transition intensities. The n values reflecting the sample variation (Aalen, Farewell, De Angelis, Day, and Gill 1997), we order them and the 2.5th and the 97.5th empirical percentiles are then used as lower and upper confidence bounds for 95% confidence intervals. This procedure can be repeated for any t, so we can obtain pointwise confidence limits for $\alpha_{hl}(\cdot)$.

5. Using SmoothHazard

5.1. How to prepare the data

Table 1 shows how the program interprets the structure of the data set. In all cases, L_i may be equal to the entry time. Some more details are necessary to distinguish the case where the ill status is known at the last follow-up time for death from the case where this is not possible.

- In case 2, if $L_i < C_i$ then it is assumed that the subject may become ill between L_i and C_i. If $L_i = C_i$ it is assumed that the subject is disease-free at time C_i. In the latter case the integral in the second term of the likelihood equals zero.

Case	Description	δ_1	δ_2	L	R	T	Remark
1	Exact ill time, right-censored death time.	1	0	L_i	L_i	C_i	$L_i \leq C_i$
2	No illness observed, right-censored death time.	0	0	L_i	L_i	C_i	$L_i \leq C_i$
3	Interval-censored ill time, right-censored death time.	1	0	L_i	R_i	C_i	$L_i < R_i \leq C_i$
4	Exact ill time, death time observed.	1	1	L_i	L_i	T_i	$L_i \leq T_i$
5	No illness observed, death time observed.	0	1	L_i	L_i	T_i	$L_i \leq T_i$
6	Interval-censored ill time, death time observed.	1	1	L_i	R_i	T_i	$L_i < R_i \leq T_i$

Table 1: Description of how the data set must be built to be understood by the `idm` function.

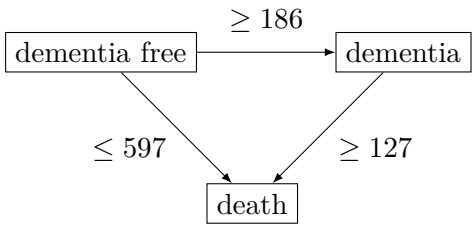

Figure 3: Example of the `Paq1000` data set: The exact number of transitions in the illness-death model with interval-censored time to disease is unknown.

- In case 5, if $L_i < T_i$ then it is assumed that the subject may become ill between L_i and T_i. If $L_i = T_i$ it is assumed that the subject is disease-free at time T_i. In the latter case the integral in the second term of the likelihood equals zero.

5.2. Paquid study

In order to illustrate the functionality of the package we provide a random subset containing data from 1000 subjects that were enrolled in the Paquid study (Letenneur, Gilleron, Commenges, Helmer, Orgogozo, and Dartigues 1999), a large cohort study on mental and physical aging.

```
R> data("Paq1000", package = "SmoothHazard")
```

The population consists of subjects aged 65 years and older living in Southwestern France. The event of interest is dementia and death without dementia is a competing risk. Furthermore, the time to dementia onset is interval-censored between the diagnostic visit and the previous one and demented subjects are at risk of death. Thus, subjects who died without being diagnosed as demented at their last visit may have become demented between last visit and death.

In this subset `186` subjects are diagnosed as demented and `724` died from whom `597` without being diagnosed as demented before. Because of interval-censoring more than `186` should have been demented, more than `127` should have been dead with dementia and less than `597` should have been dead without dementia (see Figure 3).

Age is chosen as the basic time scale and subjects are dementia-free (and alive) at entry into study. Consequently, we need to deal with left-truncated event times.

```
R> head(round(Paq1000, 1))
```

```
  dementia death    e    l    r    t certif gender
1        1     1 72.3 82.3 84.7 87.9      0      0
2        0     1 77.9 78.9 78.9 79.6      0      1
3        0     1 79.9 79.9 79.9 80.9      0      0
4        0     1 74.7 78.6 78.6 82.9      1      1
5        0     1 76.7 76.7 76.7 79.2      0      1
6        0     0 66.2 71.4 71.4 84.2      1      0
```

Each row in the data corresponds to one subject. The variables `dementia` and `death` are δ_1 and δ_2, the status variables for dementia and death. The variable `e` contains the age of subjects at entry into the study. The variables `l` and `r` contain the left and right endpoints of the censoring intervals. For demented subjects, `r` is the age at the diagnostic visit and `l` is the age at the previous one. For non-demented subjects, `l` and `r` are the age at the latest visit without dementia (`l` = `r`). The variable `t` is the age at death or at the latest news on the vital status. There are two binary covariates: `certif` for primary school diploma (`762` with diploma and `238` without diploma) and `gender` (`578` women and `422` men).

The function `idm` computes estimates for the three transition intensities $\alpha_{01}(\cdot)$, $\alpha_{02}(\cdot)$, $\alpha_{12}(\cdot)$ which represent age-specific incidence rates of dementia, age-specific mortality rate of dementia-free subjects and age-specific mortality rate of demented subjects, respectively. Proportional transition intensities regression models allow for covariates on each transition. Covariates are specified independently for the regression models of the three transition intensities by the right hand side of the respective formula `formula01`, `formula02` and `formula12`.

Interval-censoring and left-truncation must be specified at the left side of the formula arguments using the `Hist` function. For left-truncated data, the `entry` argument of `Hist` must contain the vector of delayed entry times. For interval-censored data, the `time` argument of `Hist` must contain a list of the left and right endpoints of the intervals. The `data` argument contains the data frame in which to interpret the variables of `formula01`, `formula02` and `formula12`. The left side of `formula12` argument does not need to be filled because all the data information is already contained in `formula01` and `formula02`. The left side of `formula12` argument is required only if we want the covariates impacting transition $1 \to 2$ different from those impacting transition $0 \to 2$.

5.3. Fitting the illness-death model based on interval-censored data

The main function `idm` computes estimates for the three baseline transition intensities and for the regression parameters of an illness-death model. The `method` argument by specifying the form of the transition intensities allows to select either the parametric or a semi-parametric estimation method:

- With the default value `"Weib"`, a Weibull distribution is assumed for the baseline transition intensities and the parameters are estimated by maximizing the log-likelihood.

- With the `"Splines"` value, the baseline transition intensities are approximated by linear combinations of M-splines and the parameters are estimated by maximizing the penalized log-likelihood.

We stop the iterations of the maximization algorithm when the differences between two consecutive parameter values, log-likelihood values, and gradient values is small enough. The default convergence criteria are 10^{-5}, 10^{-5} and 10^{-3}, respectively and can be changed by means of the `eps` argument.

We now illustrate how to fit the illness-death model to the `Paq1000` data set, based on interval-censored dementia times and exact death times.

In the following call, a Weibull parametrization is used for the three baseline transition intensities and we include two covariates on the transition to dementia, one covariate on the transition from no dementia to death and no covariates on the transition from dementia to death. Note that in case of missing `formula12` argument the covariates on the $1 \to 2$ transition are taken to be the same as the ones specified in the `formula02` argument.

```
R> library("SmoothHazard")
R> fit.weib <- idm(formula01 = Hist(time = list(l, r), event = dementia,
+    entry = e) ~ certif + gender, formula02 = Hist(time = t,
+    event = death, entry = e) ~ gender, formula12 = ~ 1, data = Paq1000)
R> fit.weib


Call:
idm(formula01 = Hist(time = list(l, r), event = dementia, entry = e) ~
    certif + gender, formula02 = Hist(time = t, event = death,
    entry = e) ~ gender, formula12 = ~1, data = Paq1000)

Illness-death regression model using Weibull parametrization
to estimate the baseline transition intensities.

number of subjects:  1000
number of events '0 -> 1':  186
number of events '0 -> 2' or '0 -> 1 -> 2':  724
number of covariates:  2 1 0
----
Model converged.
number of iterations:  6
convergence criteria: parameters= 7.3e-10
                    : likelihood= 2.3e-08
                    : second derivatives= 2.8e-12

                Without covariates With covariates
 log-likelihood          -3075.308       -3053.648

Parameters of the Weibull distributions: 'S(t) = exp(-(b*t)^a)'
          transition 0 -> 1 transition 0 -> 2 transition 1 -> 2
```

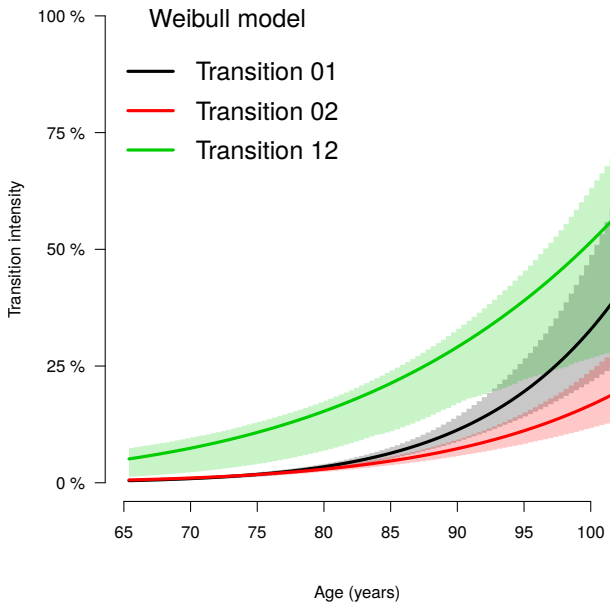

Figure 4: Baseline transition intensities estimated using the Weibull parametrization of the parametric approach.

```
shape (a)       11.12344625        8.82268159        6.44006486
scale (b)        0.01102198        0.01074539        0.01381268


Regression coefficients:

$`transition 0 -> 1`
  Factor    coef SE coef exp(coef)          CI p-value
1 certif -0.4117  0.1827    0.6625 [0.46;0.95] 0.02424
2 gender -0.2621  0.1561    0.7694 [0.57;1.04] 0.09319

$`transition 0 -> 2`
  Factor    coef SE coef exp(coef)          CI p-value
3 gender  0.6712  0.1143    1.9565 [1.56;2.45] < 1e-04
```

The hazard ratios HR (e^{coef}) have the usual interpretation, as in a parametric Cox regression model.

The three baseline transition intensity functions can be displayed as functions of time, i.e., functions of age in our illustrative example (Figure 4).

```
R> par(mgp = c(4, 1, 0), mar = c(5, 5, 5, 5))
R> plot(fit.weib, conf.int = TRUE, lwd = 3, citype = "shadow",
+    xlim = c(65, 100), axis2.las = 2, axis1.at = seq(65, 100, 5),
+    xlab = "Age (years)")
```

The other `idm` estimation option permits the relaxation of the Weibull regression model's strict parametric assumptions. With the option `method = "Splines"`, linear combinations

of M-splines are used to approximate the three baseline transition intensities. Although this option implies a considerable amount of extra computations (see Section 3.2), the call and the printed output are very similar to the Weibull model:

```
R> fit.splines <- idm(formula01 = Hist(time = list(l, r), event = dementia,
+    entry = e) ~ certif + gender, formula02 = Hist(time = t,
+    event = death, entry = e) ~ gender, formula12 = ~ 1, method = "Splines",
+    data = Paq1000)
R> fit.splines

Call:
idm(formula01 = Hist(time = list(l, r), event = dementia, entry = e) ~
    certif + gender, formula02 = Hist(time = t, event = death,
    entry = e) ~ gender, formula12 = ~1, data = Paq1000, method = "Splines")

Illness-death regression model using  M-spline approximations
to estimate the baseline transition intensities.

number of subjects:  1000
number of events '0 -> 1':  186
number of events '0 -> 2' or '0 -> 1 -> 2':  724
number of covariates:  2 1 0
----
Model converged.
number of iterations:  8
convergence criteria: parameters= 8.7e-09
                    : likelihood= 2.5e-07
                    : second derivatives= 6.9e-11

                         Without covariates With covariates
Penalized log-likelihood          -3073.099       -3052.322


Smoothing parameters:
      transition01 transition02 transition12
knots        7e+00        7e+00            7
kappa        1e+06        5e+05        20000


Regression coefficients:

$`transition 0 -> 1`
  Factor    coef SE coef exp(coef)          CI p-value
1 certif -0.3775  0.1847    0.6856 [0.48;0.98]  0.0410
2 gender -0.2376  0.1578    0.7885 [0.58;1.07]  0.1321

$`transition 0 -> 2`
```

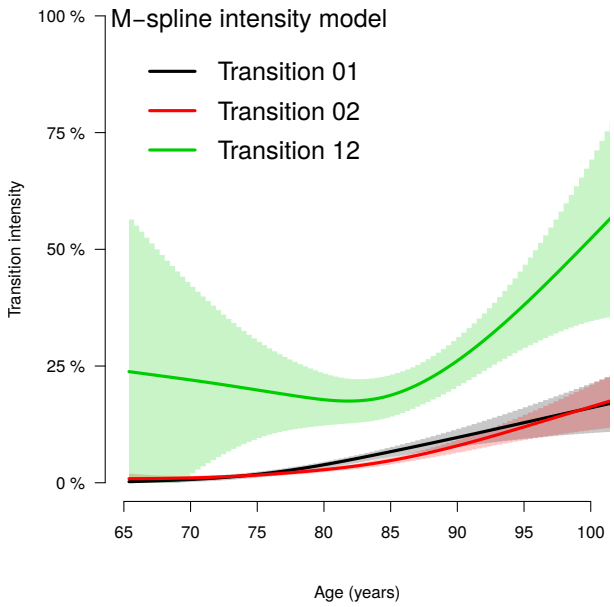

Figure 5: Baseline transition intensities estimated using the splines approximation of the semi-parametric approach.

```
  Factor    coef SE coef exp(coef)          CI p-value
3 gender  0.6628  0.1127    1.9402 [1.56;2.42]  <1e-04
```

Again, the estimated baseline transition intensities can be conveniently visualized in a joint graph (Figure 5).

```
R> par(mgp = c(4, 1, 0), mar = c(5, 5, 5, 5))
R> plot(fit.splines, conf.int = TRUE, lwd = 3, citype = "shadow",
+    xlim = c(65, 100), axis2.las = 2,  axis1.at = seq(65, 100, 5),
+    xlab = "Age (years)")
```

Figure 5 shows that the M-splines baseline transition intensities are quite different from the Weibull transition intensities. Indeed, the relaxation of the Weibull assumption permits to obtain a flexible form for the curves. At the same time, the estimation by penalized likelihood maximization yields smooth estimates.

Semi-parametric estimation method: Choice of smoothing parameters

Some optional arguments are specific to the semi-parametric approach (when using the option `method = "Splines"`):

- `n.knots` contains a vector (by default `c(7, 7, 7)`) specifying the number of knots on the $0 \to 1$, $0 \to 2$ and $1 \to 2$ transitions, respectively;
- `knots` contains the choice of the knots placement (equidistant by default or quantile-based placement) or a list of sequences of knots for transitions $0 \to 1$, $0 \to 2$ and $1 \to 2$ respectively, to be specified by the user;

- `CV` (`FALSE` by default) is set to `TRUE` for using approximate leave-one-out cross-validation score to choose the smoothing parameters κ_{01}, κ_{02}, κ_{12};
- `kappa` contains the smoothing parameters if `CV = FALSE` (arbitrary choice of the smoothing parameters κ_{01}, κ_{02}, κ_{12}); the initial smoothing parameters for the grid search method which maximize the approximate leave-one-out cross-validation score if `CV = TRUE`.

By default the function `idm` selects equidistant sequences of 7 knots between the minimal and maximal event times (`e`, `l` and `r` for `Paq1000`). There must be a knot before or at the first time from which there are subjects at risk and after or at the last time of transition. The current implementation of our program requires a minimum of 5 knots for each transition intensity.

Consequently, the semi-parametric approach requires much more information than the parametric one to achieve convergence. The number of parameters to be estimated is larger, and enough observation times on each transition are required to fit the splines. In particular, in data sets where few $1 \to 2$ transitions times are observed, we do not recommend this approach. Increasing the number of knots does not deteriorate the estimates of the transition intensities: This is because the degree of smoothing in the penalized likelihood method is tuned by the smoothing parameters κ_{01}, κ_{12} and κ_{02}. On the other hand, once a sufficient number of knots is established, there is no advantage in adding more. Moreover, the more knots, the longer the running time. Some numerical problems can arise, particularly for a large number of knots. So it is recommended to start with a small number of knots (e.g., 5 or 7) and increase the number of knots until the graph of the transition intensities function remains unchanged (from our own experience rarely more than 12 knots).

The default values for the smoothing parameters κ_{01}, κ_{02}, κ_{01}, are suitable for the `Paq1000` data set. However, these values can be expected to be very different depending on time scale, number of subjects and number of knots. The cross-validation option can be used to find appropriate smoothing parameters. However, the running time with cross-validation is very long and an empirical technique might be preferred. It consists of repeatedly running `idm` trying different smoothing parameters. After each estimation, the transition intensities are plotted. If the curves seem too smooth, it may be useful to reduce the smoothing parameter. Similarly, if the curves are too wiggly, the smoothing parameter may be increased.

5.4. Making predictions

An object returned from `idm` can be used as an argument to the `predict` function in order to obtain transition probabilities, cumulative probabilities of event and life expectancies with confidence intervals. For example, the following call gives predictions over a 10-year horizon for a 70-year-old male subject who has a primary school diploma:

```
R> pred <- predict(fit.weib, s = 70, t = 80,
+    newdata = data.frame(certif = 1, gender = 1))
R> pred

Predictions of an irreversible illness-death model with states (0,1,2).
```

```
For covariate values:

 certif gender
      1      1

For a subject in state '0' at time 70,
predicted state occupation probability at time 80:

 State Parameter Estimate Lower.95 Upper.95
     0       p00    0.635    0.590    0.681
     1       p01    0.048    0.034    0.073
     2       p02    0.317    0.272    0.354

The probability p02 can be further decomposed into
direct and indirect transition probabilities:

        Path Parameter Estimate Lower.95 Upper.95
      direct     p02_0    0.287    0.242    0.332
 via state 1     p02_1    0.030    0.012    0.043
       total       p02    0.317    0.272    0.354

For a subject in state '0' at time 70,
predicted probability of exit from state 0 until time 80:

        Path Parameter Estimate Lower.95 Upper.95
 via state 1       F01    0.078    0.049    0.116
         any       F0.    0.365    0.319    0.410

For a subject in state '1' at time 70,
predicted state occupation probability at time 80:

 State Parameter Estimate Lower.95 Upper.95
     1       p11    0.334    0.273    0.638
     2       p12    0.666    0.362    0.727
```

The output attributes are:

- for a dementia-free 70-year-old subject:
  - the probability of being still alive and dementia-free 10 years later $p_{00}(70, 80)$;
  - the probability of being still alive but demented 10 years later $p_{01}(70, 80)$;
  - the probability of dying in the next 10 years $p_{02}(70, 80)$ having been demented before ($p^1_{02}(70, 80)$) or not ($p^0_{02}(70, 80)$);
  - the absolute risk of dementia in the next 10 years (10 years later, the subject may be dead or not) $F_{01}(s, t)$;
  - the absolute risk of exit from the no dementia state in the next 10 years $F_{0\bullet}(s, t)$ (due to either dementia or death);

- for a 70-year-old demented subject:
  - the probability of dying in the next 10 years $p_{12}(70, 80)$ or not $p_{11}(70, 80)$.

The following calls give life expectancies regarding an 80-year-old female subject who has a primary school diploma based on the transition intensities estimates from respectively the parametric approach and the semi-parametric approach:

```
R> LE.weib <- predict(fit.weib, s = 80,
+    newdata = data.frame(certif = 1, gender = 0), lifeExpect = TRUE)
R> LE.weib

Predictions of an irreversible illness-death model with states (0,1,2).

For covariate values:

 certif gender
      1      0

Remaining life expected sojourn times (starting at time 80):

 State at time s Expected years in states 0,1      Parameter Estimate
               0                         Total           LE.0    8.868
               0                   In state 0 LE.nondiseased   10.364
               1                         Total    LE.diseased    4.887
 Lower.95 Upper.95
    7.910    9.854
    9.729   11.292
    4.379    7.328

R> LE.splines <- predict(fit.splines, s = 80,
+   newdata = data.frame(certif = 1, gender = 0), lifeExpect = TRUE,
+   CI = FALSE)
R> LE.splines

Predictions of an irreversible illness-death model with states (0,1,2).

For covariate values:

 certif gender
      1      0

Remaining life expected sojourn times (starting at time 80):

 State at time s Expected years in states 0,1      Parameter Estimate
               0                         Total           LE.0    8.835
               0                   In state 0 LE.nondiseased   10.437
```

```
              1                         Total    LE.diseased    5.031
Lower.95 Upper.95
   7.801    9.817
   9.609   11.136
   4.245    5.888
```

The output attributes of the `predict` function with `lifeExpect = TRUE` are:

- for an 80-year-old dementia-free subject:
  - the life expectancy in state 0 (healthy life expectancy);
  - the life expectancy;
- for an 80-year-old demented subject:
  - the life expectancy.

The confidence intervals calculation using the simulation-based method may take time, especially using the splines estimates of the transition intensities. To suppress this calculation, the `CI` argument must be set to `FALSE` (see above). Note that to reduce the computation time of the confidence intervals, the number of simulations is 200 by default but, to improve precision, it can be modified using the `nsim` argument.

6. Discussion

In this article, we have described the methods implemented in the R package **SmoothHazard** for fitting illness-death models to interval-censored transition times to the intermediate state and exact transition times to the absorbing state. Note that the package also fits simple survival models (two-state models) and also models where the transition times are right-censored but not interval-censored.

We have also explained and illustrated the actual use of **SmoothHazard** and note that several extensions are in the development phase. One extension is a model which assumes equality or proportionality between the two transition intensities α_{02} and α_{12}. Another is a model which assumes the same covariate effects for these transition intensities. At the moment, the illness-death model implemented in the package assumes a Markov process. However, several implementations are currently tested of semi-Markov models which allow that the transition intensity α_{12} depends on both the current time and the time spent in the intermediate state. We also plan to implement regression models for interval-censored observations of other multi-state models, for example the progressive disease model. The development version of the package is available at https://github.com/tagteam/SmoothHazard.

Acknowledgments

This work was supported by the grant 2010 PRSP 006 01 from the *Agence Nationale de la Recherche* for the MOBIDYQ project and by the *Région Aquitaine*.

References

Aalen OO, Farewell VT, De Angelis D, Day NE, Gill ON (1997). "A Markov Model for HIV Disease Progression Including the Effect of HIV Diagnosis and Treatment: Application to AIDS Prediction in England and Wales." *Statistics in Medicine*, **16**(19), 2191–2210. doi:10.1002/(sici)1097-0258(19971015)16:19<2191::aid-sim645>3.3.co;2-x.

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, Keiding N (2012). "Interpretability and Importance of Functionals in Competing Risks and Multistate Models." *Statistics in Medicine*, **31**(11–12), 1074–1088. doi:10.1002/sim.4385.

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.

Boruvka A, Cook RJ (2015). ***coxinterval**: Cox-Type Models for Interval-Censored Data*. R package version 1.2, URL https://CRAN.R-project.org/package=coxinterval.

Commenges D, Joly P, Gégout-Petit A, Liquet B (2007). "Choice between Semi-Parametric Estimators of Markov and Non-Markov Multi-State Models From Coarsened Observations." *Scandinavian Journal of Statistics*, **34**(1), 33–52. doi:10.1111/j.1467-9469.2006.00536.x.

Cox DR (1975). "Partial Likelihood." *Biometrika*, **62**(2), 269–276. doi:10.1093/biomet/62.2.269.

de Wreede LC, Fiocco M, Putter H (2010). "The **mstate** Package for Estimation and Prediction in Non- and Semi-Parametric Multi-State and Competing Risks Models." *Computer Methods and Programs in Biomedicine*, **99**(3), 261–274. doi:10.1016/j.cmpb.2010.01.001.

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.

Delord M (2016). ***MIICD**: Multiple Imputation for Interval Censored Data*. R package version 2.3, URL https://CRAN.R-project.org/package=MIICD.

Fay MP, Shaw PA (2010). "Exact and Asymptotic Weighted Logrank Tests for Interval Censored Data: The **interval** R Package." *Journal of Statistical Software*, **36**(2). doi:10.18637/jss.v036.i02.

Gandrud C (2015). "**simPH**: An R Package for Illustrating Estimates from Cox Proportional Hazard Models Including for Interactive and Nonlinear Effects." *Journal of Statistical Software*, **65**(3), 1–20. doi:10.18637/jss.v065.i03.

Harrell FE (2017). *Regression Modeling Strategies*. R package version 5.1-1, URL https://CRAN.R-project.org/package=rms.

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.

Joly P, Commenges D, Helmer C, Letenneur L (2002). "A Penalized Likelihood Approach for an Illness-Death Model with Interval-Censored Data: Application to Age-Specific Incidence of Dementia." *Biostatistics*, **3**(3), 433–443. doi:10.1093/biostatistics/3.3.433.

Keiding N (2014). "Event History Analysis." *Annual Review of Statistics and Its Application*, **1**, 333–360. doi:10.1146/annurev-statistics-022513-115558.

Leffondré K, Touraine C, Helmer C, Joly P (2013). "Interval-Censored Time-to-Event and Competing Risk with Death: Is the Illness-Death Model More Accurate Than the Cox Model?" *International Journal of Epidemiology*, **42**(4), 1177–1186. doi:10.1093/ije/dyt126.

Letenneur L, Gilleron V, Commenges D, Helmer C, Orgogozo JM, Dartigues JF (1999). "Are Sex and Educational Level Independent Predictors of Dementia and Alzheimer's Disease? Incidence Data from the PAQUID Project." *Journal of Neurology, Neurosurgery & Psychiatry*, **66**(2), 177–183. doi:10.1136/jnnp.66.2.177.

Levenberg K (1944). "A Method for the Solution of Certain Problems in Least Squares." *Quarterly of Applied Mathematics*, **2**, 164–168. doi:10.1090/qam/10666.

Mandel M (2013). "Simulation Based Confidence Intervals for Functions with Complicated Derivatives." *The American Statistician*, **67**(2), 76–81. doi:10.1080/00031305.2013.783880.

Marquardt DW (1963). "An Algorithm for Least-Squares Estimation of Nonlinear Parameters." *Journal of the Society for Industrial & Applied Mathematics*, **11**(3), 431–441. doi:10.1137/0111030.

McMahan CS, Wang L (2014). ***ICsurv**: A Package for Semiparametric Regression Analysis of Interval-Censored Data*. R package version 1.0, URL https://CRAN.R-project.org/package=ICsurv.

Meira-Machado L, Roca-Pardiñas J (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.

O'Sullivan F (1988). "Fast Computation of Fully Automated Log-Density and Log-Hazard Estimators." *SIAM Journal on Scientific and Statistical Computing*, **9**(2), 363–379. doi:10.1137/0909024.

Pan W (1999). "Extending the Iterative Convex Minorant Algorithm to the Cox Model for Interval-Censored Data." *Journal of Computational and Graphical Statistics*, **8**(1), 109–120. doi:10.1080/10618600.1999.10474804.

Putter H, Fiocco M, Geskus RB (2007). "Tutorial in Biostatistics: Competing Risks and Multi-State Models." *Statistics in Medicine*, **26**(11), 2389–2430. doi:10.1002/sim.2712.

Ramsay JO (1988). "Monotone Regression Splines in Action." *Statistical Science*, **3**(4), 425–441. doi:10.1214/ss/1177012761.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reich NG, Lessler J, Azman AS (2016). ***coarseDataTools***: *A Collection of Functions to Help with Analysis of Coarsely Observed Data*. R package version 0.6-3, URL https://CRAN.R-project.org/package=coarseDataTools.

Reich NG, Lessler J, Cummings DAT, Brookmeyer R (2009). "Estimating Incubation Periods with Coarse Data." *Statistics in Medicine*, **28**(22), 2769–2784. doi:10.1002/sim.3659.

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

Therneau TM, Grambsch PM (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York. doi:10.1007/978-1-4757-3294-8.

Touraine C, Helmer C, Joly P (2016). "Predictions in an Illness-Death Model." *Statistical Methods in Medical Research*, **25**(4), 1452–1470. doi:10.1177/0962280213489234.

Touraine C, Joly P, Gerds TA (2017). ***SmoothHazard***: *Estimation of Smooth Hazard Models for Interval-Censored Data with Applications to Survival and Illness-Death Models*. R package version 1.4.0, URL https://CRAN.R-project.org/package=SmoothHazard.

Affiliation:

Célia Touraine
University of Bordeaux
ISPED
Inserm Research Centre U1219
Bordeaux F-33000, France
E-mail: celia.touraine@icm.unicancer.fr
URL: http://www.bordeaux-population-health.center/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
July 2017, Volume 79, Issue 7 *Submitted:* 2014-12-29
doi:10.18637/jss.v079.i07 *Accepted:* 2016-06-30