An R package for Inference and Prediction in an Illness-Death Model

Luís Meira-Machado[1], Marta Sestelo[2]

[1] Centre of Molecular and Environmental Biology & Department of Mathematics and Applications, University of Minho, Campus de Azurem, 4800-058 Guimarães, Portugal.
[2] SiDOR Research Group and CINBIO, University of Vigo, Spain.

E-mail for correspondence: lmachado@math.uminho.pt

Abstract: Multi-state models are a useful way of describing a process in which an individual moves through a number of finite states in continuous time. The illness-death model plays a central role in the theory and practice of these models, describing the dynamics of healthy subjects who may move to an intermediate 'diseased' state before entering into a terminal absorbing state. In these models one important goal is the modeling of transition rates which is usually done by studying the relationship between covariates and disease evolution. However, biomedical researchers are also interested in reporting other interpretable results in a simple and summarized manner. These include estimates of predictive probabilities, such as the transition probabilities, occupation probabilities, cumulative incidence functions, prevalence and the sojourn time distributions. An **R** package was built providing answers to all these topics.

Keywords: Illness-death model; Kaplan-Meier; Landmark approach; Nonparametric estimation; Survival analysis.

1 Introduction

Multi-state models are very useful for describing complex event history data. These models may be considered a generalization of survival analysis where survival is the ultimate outcome of interest but where information is available about intermediate events which individuals may experience during the study period. For instances, in most biomedical applications, besides the 'healthy' initial state and the absorbing 'dead' state, one may observe intermediate (transient) states based on health conditions, disease stages, clinical symptoms, etc. The illness-death model is probably the most

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018.

popular one in the medical literature. The irreversible version of this model (Figure 1), describes the pathway from an initial state to an absorbing state either directly or through an intermediate state. Many time-to-event data sets from biomedical studies with multiple events can be reduced to this generic structure. Recent reviews on this topic may be found in the papers by Putter et al. (2007), Meira-Machado et al. (2009), Meira-Machado et al. (2011) and Meira-Machado and Sestelo (2018).

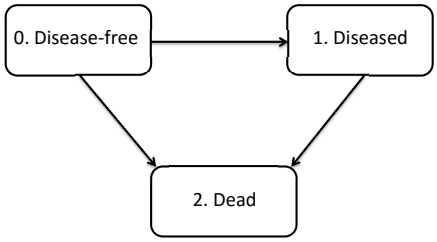

FIGURE 1. Illness-death model.

One important goal in multi-state modelling is to relate the individual characteristics with the intensity rates through a covariate vector but biomedical researchers are also interested in reporting interpretable results in a simple and summarized manner. These include estimates of predictive probabilities, such as the transition probabilities, occupation probabilities, cumulative incidence functions, prevalence and the sojourn time distributions. The development of **survidm R** package has been motivated by several recent contributions that account for these problems; in particular the newly developed methods based on landmarking. The current version of the package provides seven different approaches to estimate the transition probabilities, three methods for the sojourn distributions and one approach for the cumulative incidence functions. In addition, these probabilities can also be estimated conditionally on covariate measures. The package also allows the user to perform multi-state regression where the estimation of the covariate effects is achieved using Cox regression in which different effects of the covariates are assumed for different transitions.

2 survidm in practice

This software enables both numerical and graphical outputs to be displayed for several methods. This software is intended to be used with the R statistical program. Our package is composed of 13 functions that allow users to obtain estimates for all proposed methods. Details on the usage of the functions (described in Table 1) can be obtained with the corresponding help pages.

Function	Description
`survIDM`	Create a `survIDM` object.
`coxidm`	Fits proportional hazards regression models for each transition.
`tprob`	Nonparametric estimation of the transition probabilities.
`CIF`	Nonparametric estimation of the cumulative incidence functions.
`sojourn`	Nonparametric estimation of the sojourn distributions.
`plot.survIDM`	Plot for an object of class `survIDM`.
`print.survIDM`	Print for an object of class `survIDM`.
`summary.survIDM`	Summary for an object of class `survIDM`.
`KM`	Computes the Kaplan-Meier product-limit of survival.
`PKM`	Computes the presmoothed Kaplan-Meier product-limit of survival.
`Beran`	Computes the conditional survival probability of the response, given the covariate under random censoring.
`KMW`	Returns a vector with the Kaplan-Meier weights.
`PKMW`	Returns a vector with the presmoothed Kaplan-Meier weights.
`LLW`	Returns a vector with the local linear weights.
`NWW`	Returns a vector with the Nadaraya-Watson weights.

TABLE 1. Summary of functions in the **survidm** package.

It should be noted that to implement the methods described in the methodology section one needs the following variables of data: `time1`, `event1`, `Stime` and `event`. A single covariate can also be included (it is only necessary for `IPCW` methods). The variable `time1` represents the observed time to the first event of interest, and `event1` the corresponding status/censoring indicator (if the survival time is a censored observation, the value is 0 and otherwise the value is 1). The variable `Stime` represents the total survival time. If `event1` $= 0$, then the total survival time is equal to the observed time to the first event. The variable `event` is the final status of the individual (takes the value 1 if the final event of interest is observed and 0 otherwise).

For illustration purposes we will use data of 929 patients affected by colon cancer that underwent a curative surgery for colorectal cancer. In this study, 468 developed recurrence and among these 414 died. 38 patients died without recurrence. The rest of the patients (423) remained alive and disease-free up to the end of the follow-up. Besides the two event times (time to recurrence and time to death) and the corresponding indicator statuses a vector of covariates including `age`, `sex` and number of lymph nodes (`nodes`) are also available.

One important goal in multi-state modeling is to study the relationships

between the different predictors and the outcome. To relate the individual characteristics to the intensity rates several models have been used in literature. A common simplifying strategy is to decouple the whole process into various survival models, by fitting separate intensities to all permitted transitions using semi-parametric Cox proportional hazard regression models, while making appropriate adjustments to the risk set. This can be obtained using the following input commands:

```
library(survidm)
data(colonIDM)
fit.cmm <- coxidm(survIDM(time1, event1, Stime, event) ~ age
       + sex + nodes, data = colonIDM)
summary(fit.cmm)
```

Results obtained from the above input commands (not shown) reveal that multi-state regression models provide detailed information of the disease process, revealing how the different covariates may affect the various permitted transitions. For instances, it revealed `age` as an important predictor on the mortality transitions (with and without recurrence) but not on the recurrence incidence, whereas `sex` only revealed a significant effect on the mortality transition after recurrence.
The patients course over time may also be studied through other quantities such as the transition probabilities. To obtain these estimates (for a model with no covariates), the following input command must be typed:

```
res <- tprob(survIDM(time1, event1, Stime, event) ~ 1, s=365,
       method = "LM", conf=TRUE, data = colonIDM)
summary(res, time=365*1:6)
plot(res)
```

Figure 2 reports estimated transition probabilities ($P_{ij}(s,t)$) for a fixed value of $s = 365$ (days), along time. Results were obtained using the Landmark method (`method = "LM"`) proposed by de Uña-Álvarez and Meira-Machado (2015). It is worth mention that function `tprob` implements eight distinct methods including the possibility of estimating these quantities conditional on covariates.
Estimates and plots for the cumulative incidence (of recurrence) (Geskus 2011) and for the sojourn time distribution quantities can also be obtained. The following input commands provide the corresponding numerical and graphical output for the two quantities:

```
res.cif <- CIF(survIDM(time1, event1, Stime, event) ~ 1,
           data = colonIDM, conf = TRUE)
summary(res.cif, time = 365*1:7)
plot(res.cif, ylim=c(0, 0.6))
```

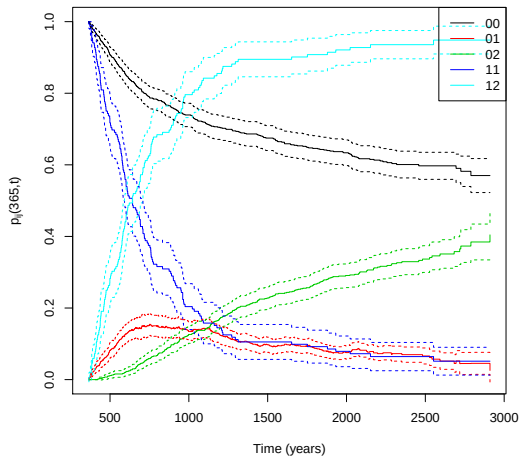

FIGURE 2. Estimates of the transition probabilities using the landmark method. Colon cancer data.

```
res.soj <- sojourn(survIDM(time1, event1, Stime, event) ~ 1,
           data = colonIDM, conf = TRUE, conf.level = 0.95)
summary(res.soj, time = 365*1:6)
plot(res.soj)
```

References

Moreira, A., de Uña-Álvarez, J. and Meira-Machado, L. (2013). Presmoothing the Aalen-Johansen estimator in the illness-death model. *Electronic Journal of Statistics*, **7**, 1491 – 1516.

Meira-Machado, L., de Uña-Álvarez, J. and Somnath, D. (2015). Conditional Transition Probabilities in a non-Markov Illness-death Model. *Computational Statistics*, **30(2)**, 377 – 397.

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

Putter, H. and Spitoni, C. (2016). Non-parametric estimation of transition probabilities in non-Markov multi-state models: The landmark Aalen-Johansen estimator. *Statistical Methods in Medical Research*, 1 – 12.

Geskus, R.B. (2011). Cause-Specific Cumulative Incidence Estimation and the Fine and Gray Model Under Both Left Truncation and Right Censoring. *Biometrics*, **67**, 39 – 49.

Meira-Machado, L. (2011). Inference for non-Markov multi-state models: an overview. *REVSTAT - Statistical Journal*, **9 (1)**, 83 – 98.

Meira-Machado, L. (2016). Smoothed landmark estimators of the transition probabilities. *SORT-Statistics and Operations Research Transactions*, **40**, 375 – 398.

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

Meira-Machado, L., Sestelo, M. (2018). Estimation in the progressive illness-death model: a non-exhaustive review. *submitted*.