Published in final edited form as:

Stat Methods Med Res. 2009 April ; 18(2): 195–222. doi:10.1177/0962280208092301.

Multi-state models for the analysis of time-to-event data

Luís Meira-Machado

Department of Mathematics for Science and Technology, University of Minho, Portugal

Jacobo de Uña-Álvarez

Department of Statistics and Operations Research, University of Vigo, Spain

Carmen Cadarso-Suárez

Department of Statistics and Operations Research, University of Santiago de Compostela, Spain

Per K Andersen

Department of Biostatistics, University of Copenhagen, Denmark

Abstract

The experience of a patient in a survival study may be modelled as a process with two states and one possible transition from an “alive” state to a “dead” state. In some studies, however, the “alive” state may be partitioned into two or more intermediate (transient) states, each of which corresponding to a particular stage of the illness. In such studies, multi-state models can be used to model the movement of patients among the various states. In these models issues, of interest include the estimation of progression rates, assessing the effects of individual risk factors, survival rates or prognostic forecasting. In this article, we review modelling approaches for multi-state models, and we focus on the estimation of quantities such as the transition probabilities and survival probabilities. Differences between these approaches are discussed, focussing on possible advantages and disadvantages for each method. We also review the existing software currently available to fit the various models and present new software developed in the form of an R library to analyse such models. Different approaches and software are illustrated using data from the Stanford heart transplant study and data from a study on breast cancer conducted in Galicia, Spain.

1 Introduction

In longitudinal medical studies, patients are observed over time and covariate information is collected at several occasions. The analysis in such studies where individuals may experience several events is often performed using multi-state models. A multi-state model (MSM) is a model for a continuous time stochastic process allowing individuals to move among a finite number of states. In biomedical applications, the states might be based on clinical symptoms (e.g. bleeding episodes), biological markers (e.g. CD4 T-lymphocyte cell counts; serum immunoglobulin levels), some scale of the disease (e.g. stages of cancer or HIV infection) or a non-fatal complication in the course of the illness (e.g. cancer recurrence). A change of state is called a transition, or an event. States can be transient or absorbing, if no transitions can emerge from the state (for example, death). The complexity of a MSM greatly depends on the number of states and also on the possible transitions.

The simplest form of an MSM is the *mortality model* for survival analysis with states “alive” and “dead” and only one possible transition. Splitting the “alive” state into two transient states, we obtain the simplest *progressive three-state model*. Both models are special cases of the *k-progressive model*, illustrated in Figure 1.

Another possible MSM widely used in the medical literature to describe the disease progression is the *illness-death model* (also known as disability model, Figure 2) which can be used to study the incidence of the disease and the rate of death. One problem here may be to evaluate whether previously diseased subjects have the same rate of death as those who have been healthy all their lives.

The *competing risks model*¹ is another MSM which extends the simple mortality model for survival data in which each individual may “die” due to any of several causes. The *bivariate model*,² depicted in Figure 3, is the MSM for bivariate parallel data, with states “both alive”, “individual 1 dead”, “individual 2 dead” and “both dead”.

In an MSM, the transition intensities provide the hazards for movement from one state to another. These functions can also be used to determine the mean sojourn time in a given state and the number of individuals in different states at a certain moment. Covariates may be incorporated in models through transition intensities to explain differences among individuals in the course of the illness. Often, intermediate events change the natural history of the disease progression so that the role of some of the prognostic factors may not be the same after that event and also, prognostic factors associated with the rate of death can be different in different transient states.

In medical applications, observation of the process will often be ended before an absorbing state is reached leading to right-censored observation times. It can also happen that the process is not observed from its origin, leading to left-censored observations if events are known to have occurred before entry but the times of these events are unknown. Frequently, patients are only observed at intermittent follow-up visits in which case complete information from the periods between visits is not available, e.g. the transition times are not exactly observed and the states occupied in between visits are not known; these observations are said to be interval-censored. Also, left-truncation may be present when a process is not observed from its origin but only conditionally on having survived until a later point in time. The presence of these features creates special problems in analysing the data and needs to be carefully considered when constructing likelihood functions. For the examples considered in Section 4, event times are observed exactly or they are right-censored, though a likelihood construction allowing interval-censored observations is given when reviewing some multi-state approaches (Section 2). Details about how to treat incomplete observations can be found in the book by Klein and Moeschberger³ or in paper by Commenges.⁴

In most cases, the truncation and censoring mechanisms are assumed to be independent from the process. Independent right-censoring (loosely) means that the censored individuals can be represented by those without censoring.

There exists an extensive literature on MSMs. Main contributions include books by Andersen *et al.*⁵ and Hougaard.² Recent reviews on this topic may be found in the papers by Commenges,⁶ Hougaard⁷ and Andersen and Keiding.⁸ An issue of the journal *Statistical Methods in Medical Research*, entirely devoted to these models, was published in 2002. However, we believe that multi-state models have a potential to being used much more by practitioners, and that the lack of knowledge of the available software may be responsible for this lack of popularity. The aim of this article is, therefore, to review the most commonly used MSMs in an applied framework (Sections 2 and 4), and to report on the existing software (including a routine developed by the authors, Section 3). In Section 4, two data sets are analysed to illustrate

the methods and the software. One of these data sets (the Galician breast cancer data) is new but we also use the widely studied Stanford heart transplantation database for illustration.

The article is organised as follows. Sections 2 focusses on methodological aspects emphasising the estimation of the effects of covariates and estimation of transition probabilities. In this section, we also review some recent proposals for the estimation of transition probabilities which do not rely on the Markov assumption. In Section 3, we present software for implementing multi-state models and we briefly describe our R program. Results obtained from fitting the different models to the mentioned applications are compared in Section 4, in which the authors' R program output is fully illustrated. Finally some discussion is given in Section 5.

2 Multi-state models

A multi-state process is a stochastic process $(X(t), t \in T)$ with a finite state space $S = \{1, \dots, N\}$. Here, $T = [0, \tau]$, $\tau < \infty$ is a time interval and the value of the process at time t the state occupied at that time. With the evolution of the process over time, a history H_{t-} (a σ -algebra, will be generated consisting of the observation of the process over the interval $[0, t)$, such as the states previously visited, times of transitions, etc. The multi-state process is fully characterised through transition probabilities between states h and j ,

$$p_{hj}(s, t) = \mathbb{P}(X(t) = j | X(s) = h, H_{s-})$$

for $h, j \in S$, $s, t \in T$, $s \leq t$ or through transition intensities

$$\alpha_{hi}(t) = \lim_{\Delta t \rightarrow 0} \frac{p_{hi}(t, t + \Delta t) - p_{hi}(t, t)}{\Delta t} \quad (1)$$

representing the instantaneous hazard of progression to state j conditionally on occupying state h , and that we shall assume exist. Notice that both $p_{hj}(s, t)$ and $\alpha_{hj}(t)$ depend on the history though we suppress this in the notation.

Different model assumptions can be made about the dependence of the transition rates (1) on time. These include:

1. Time homogeneous models: the intensities are constant over time, that is, independent of t .^{9,10}
2. Markov models: the transition intensities only depend on the history of the process through the current state.⁹⁻¹¹
3. Semi-Markov models: future evolution not only depends on the current state h , but also on the entry time t_h into state h . Therefore, we may consider intensity functions of the general form $\alpha_{hj}(t, t - t_h)$ or, as the special homogeneous case $\alpha_{hj}(t - t_h)$.¹²

2.1 Markov models

Because of their simplicity, Markov models have been most frequently used. The process $(X(t), t \geq 0)$ is Markovian, if for any s, t with $0 \leq s < t$, we have

$$\mathbb{P}(X(t) = j | X(s) = h, H_{s-}) = \mathbb{P}(X(t) = j | X(s) = h),$$

thus, the future of the process after time s depends only on the state occupied at time s .

In Markov models, the transition probabilities can be calculated from the intensities by solving the so-called forward Kolmogorov differential equation.¹¹ For example, for the illness-death model the transition probabilities have explicit expression,

$$p_{11}(s, t) = e^{-(A_{12}(s, t) + A_{13}(s, t))}, \quad (2)$$

$$p_{22}(s, t) = e^{-A_{23}(s, t)}, \quad (3)$$

$$p_{12}(s, t) = \int_s^t p_{11}(s, u) \alpha_{12}(u) p_{22}(u, t) du, \quad (4)$$

where $A_{hj}(s, t) = \int_s^t \alpha_{hj}(u) du$ is the cumulative or integrated intensity between s and t .

We shall now review some special multi-state Markov models.

2.1.1 Time-homogeneous model—In time homogeneous Markov models (THMM), all transition intensities are assumed to be constant as functions of time. Therefore, each transition probability, $p_{hj}(s, t)$, depends only on $t - s$, that is, $p_{hj}(s, t) = p_{hj}(0, t - s)$. To simplify notation, we will use only one argument in time $p_{hj}(t - s) = p_{hj}(0, t - s)$. Then, the Kolmogorov differential equation has an explicit solution using the decomposition of the intensity matrix into eigenvalues and eigenvectors. See, for example, Cox and Miller.¹¹

Solutions for Equations (2)–(4) are now the following:

$$p_{11}(t) = e^{-(\alpha_{12} + \alpha_{13})t}, \quad p_{12}(t) = \frac{\alpha_{12}(e^{-\alpha_{23}t} - e^{-(\alpha_{12} + \alpha_{13})t})}{\alpha_{12} + \alpha_{13} - \alpha_{23}} \text{ and } p_{22}(t) = e^{-\alpha_{23}t}.$$

Inference in these may be based on the likelihood presented in Section 2.1.3.

2.1.2 Non-homogeneous models—In some applications, the hypothesis of homogeneity may be unrealistic and a nonhomogeneous model needed. A simple (parametric) procedure consists of partitioning time into two or more intervals and to assume a piecewise constant intensity model.^{13,14} Here, the transition intensities are defined by

$$\alpha_{hj}(t) = \alpha_{hj}^l, \quad \theta_{l-1} < t \leq \theta_l, \quad l = 1, 2, 3, \dots, q$$

where $\theta = (\theta_1, \dots, \theta_{q-1})$ is the vector of cutpoints ($0 = \theta_0 < \theta_1 < \theta_2 < \dots < \theta_{q-1} < \theta_q = \infty$).

Non-homogeneous Markov processes may also be modelled non-parametrically. This can be thought of as the generalisation of the Kaplan–Meier approach for the simple mortality model and was proposed by Aalen and Johansen¹⁵ for general MSMs with a finite number of states. Other non-parametric approaches to non-homogeneous processes are also available.¹⁶⁻¹⁷

2.1.3 Inference—The inference for the time-homogeneous model is conducted by a likelihood, which is derived as follows. Assume first that the stochastic process $X(\cdot)$ is observed

at times $t_{i,0} < t_{i,1} < \dots < t_{i,m_i}$, where $i = 1, \dots, n$ index individuals. That is, we observe the states $x_{i,r} = X_i(t_{i,r})$ occupied at these time points and thereby allow for interval-censoring. The likelihood is then the product over all individuals and all the time points,¹⁰

$$L = \prod_{i=1}^n \prod_{r=0}^{m_i-1} l_{i,r} \quad (5)$$

where $l_{i,r} = p_{x_{i,r}, x_{i,r+1}}(t_{i,r+1} - t_{i,r})$ is the transition probability for the i th individual corresponding to the states observed.

The likelihood function for the piecewise constant model may be built in a similar way though the expressions for $l_{i,r}$ become more complicated.¹³

In the special case where the exact transition times are all known explicit estimators for all α_{hj}^l are available. These are given by $\hat{\alpha}_{hj}^l = (m_{hj}^l) / (T_h^l)$ where m_{hj}^l is the total number of observed $h \rightarrow j$ transitions in interval number l and T_h^l is the total observation time spent in state h during the interval number l .

The transition probabilities (2) and (3) may also be estimated using Kaplan–Meier estimators via the non-parametric (Aalen–Johansen estimator) model:

$$\widehat{p}_{11}(s, t) = \prod_{s < t_{(k)} \leq t} \left(1 - \frac{d_{12k} + d_{13k}}{n_{1k}} \right) \quad (6)$$

$$\widehat{p}_{22}(t) = \prod_{s < t_{(k)} \leq t} \left(1 - \frac{d_{23k}}{n_{2k}} \right) \quad (7)$$

where $t_{(1)} < t_{(2)} < \dots < t_{(d)}$ are the event times (for disease and death) arranged in increased order, n_{1k} and n_{2k} denote the number of healthy and diseased subjects, respectively, just prior to the event time $t_{(k)}$. Further, d_{12k} is the number of subjects who become diseased at time $t_{(k)}$, while d_{13k} and d_{23k} denote, respectively, the numbers of healthy and diseased subjects who die at that same time.

An estimator for (4) is given by

$$\widehat{p}_{12}(s, t) = \sum_{s < t_{(k)} \leq t} \widehat{p}_{11}(s, t_{(k-1)}) \frac{d_{13k}}{n_{1k}} \widehat{p}_{22}(t_{(k)}, t) \quad (8)$$

which is a plug-in estimator, obtained from (4) by replacing $p_{11}(s, u) = p_{11}(s, u-)$ by $\widehat{p}_{11}(s, u-)$, $p_{22}(u, t)$, $p_{22}(u, t)$ by $\widehat{p}_{22}(u, t)$ and $\alpha_{12}(u)$ by $d\widehat{A}_{12}(u)$ the increment of the Nelson–Aalen estimator $\widehat{A}_{12}(u) = \sum_{t_k \leq u} d_{12k} / n_{1k}$ for the cumulative disease intensity $A_{12}(t) = \int_0^t \alpha_{12}(u) du$.

Because (6) and (7) are Kaplan–Meier estimators, we may use Greenwood's formula to achieve a variance estimator, whereas, a variance estimator for (8) can be found in Andersen *et al.*⁵ and Keiding *et al.*¹⁸ where general cases can also be found.

The assumption of time homogeneity may be assessed using a piecewise constant model. A likelihood ratio test can be used to compare the piecewise constant model with the time homogeneous model. Under the null hypothesis of homogeneity, the test statistic has approximately a χ^2_{q-r} distribution, where r is the number of parameters under H_0 and q the number of parameters under H_1 . Alternatively, the local score test can be used. A complete description of this method can be found in Kalbfleisch and Lawless.¹⁹

2.2 Non-Markov models

There are few references to estimation in non-Markov MSMs in the literature, some exceptions being the works of Pepe,²⁰ Strauss and Shavelle,²¹ Aalen *et al.*,²² Datta and Satten,²³ Glidden²⁴ and Meira-Machado *et al.*²⁵. Pepe²⁰ proposed a non-parametric estimator for the state occupation probability in the illness-death model based on the differences between Kaplan–Meier estimators. Strauss and Shavelle²¹ developed an extension of the Kaplan–Meier estimator for the estimation of transition probabilities, avoiding the Markov assumption. The proposed estimator is constructed by partitioning the survival probability in proportion to the number of alive and uncensored patients in each state. Aalen *et al.*²² and Datta and Satten²³ studied the performance of the Aalen–Johansen estimator of stage occupation probabilities when the process is not Markovian. These authors established the consistency of Aalen–Johansen estimators of these probabilities in a non-Markov process under independent censoring. Later, Glidden²⁴ developed robust confidence bands for those event curves. Recently, Meira-Machado *et al.*²⁵, showed that, in non-Markov situations, the use of Aalen–Johansen estimators to estimate the transition probabilities, $p_{hj}(s, t)$, may be inappropriate. These authors proposed estimators for the transition probabilities which do not rely on the Markov assumption. The new estimators are motivated as natural extensions to a censored scenario of the proportion of individuals in each state.

The ideas behind the estimators in Meira-Machado *et al.*²⁵ can be applied to introduce empirical transition probabilities in other non-Markov MSMs such as the recurrent events or the bivariate model. An issue of much practical interest is that of testing the Markov assumption. Usually, this assumption is checked including covariates depending on the history.^{10,12} However, since the methods in Meira-Machado *et al.*²⁵ are free of the Markov assumption, they can also be used to introduce such tests (at least in the scope of the illness-death model) by measuring their discrepancy to Markovian estimators. This topic is currently under investigation.

2.3 Multi-state regression models

Defining $X_i(t)$ as in (1), we are assuming that the intensities are the same for all subjects. In practical situations, however, it might be valuable to relate the individual characteristics to the intensity rates through a covariate vector, Z , possibly time-dependent. For a general regression model we can write

$$\alpha_{hji}(\cdot) = \varphi(\alpha_{hj,0}(\cdot), \beta_{hj}^T Z_i)$$

where $\alpha_{hj,0}(\cdot)$ is the baseline intensity function between states h and j , β_{hj} is the vector of regression parameters, and Z_i is the covariate vector for subject i .

A popular choice is the proportional hazards model, which is obtained by choosing $\varphi(u(\cdot), v) = u(\cdot) e^v$, that is,

$$\alpha_{hj}(t;Z) = \alpha_{hj,0}(t) \exp(\beta_{hj}^T Z) \quad (9)$$

An alternative allowing for time-dependent regression coefficients $\beta_{hj}(t)$ was proposed by Aalen for survival data and later used in multi-state models. This is obtained by choosing $\varphi(u(\cdot), v) = u(\cdot) + v$, that is,

$$\alpha_{hj}(t;Z) = \alpha_{hj,0}(t) + \beta_{hj}^T(t) Z. \quad (10)$$

For THMM and for the piecewise homogeneous model (NHM) a Cox proportional hazards model of type

$$\alpha_{hj}^l(Z) = \alpha_{hj}^l \exp(\beta_{hj}^T Z) \quad (11)$$

is traditionally assumed to relate the transition intensities α_{hj}^l in the l th time interval to covariates Z . The introduction of covariates in these models allows prediction of probabilities tailored to individual patients. In both the homogenous and piecewise homogeneous cases the inference may be based on the likelihood (5) by replacing the transition intensities there by those given above in Equation (11).

2.3.1 Cox-like transition intensities—The inference problem in a MSM can be decoupled into various survival models, by fitting separate intensities to all permitted transitions. For simplicity, assume the illness-death model of Figure 2. The transition intensities, $\alpha_{hj}(t; Z)$, $1 \leq h < j \leq 3$, may be modelled using Cox-like models of the form (9) assuming the process to be *Markovian*. These models are known as Cox Markov models (CMM) and without covariates they reduce to the non-parametric models of Section 2.1.2.

For the hazard of “death” without disease, $\alpha_{13}(t; Z)$, survival times from patients that experienced the disease are treated as censored at the observation disease time. Patients that are alive and disease-free also contribute with censored survival times. For the disease intensity, $\alpha_{12}(t; Z)$, the final point is the time of disease. Survival times of patients who did not become diseased are treated as censored, whether they are alive or whether they have died without having been diseased. Finally, to model $\alpha_{23}(t; Z)$, the death intensity after the occurrence of the disease, we only enter the survival times (censored or not) truncated on disease time of the individuals that experienced the disease. Note that patients are at risk only after entering state 2.

In some cases we may impose some restriction on the baseline hazards. For example, for the illness-death model, one approach that is often considered is to assume the baseline hazards for transition $1 \rightarrow 3$ and for the $2 \rightarrow 3$ transition to be proportional. In such cases, the model for these transitions is given by

$$\alpha_{13}(t;Z) = \alpha_{13,0}(t) \exp(\beta_{13}^T Z) \text{ and } \alpha_{23}(t;Z) = \alpha_{13,0}(t) \exp(\beta_{23}^T Z + \delta).$$

.

The estimation of the transition probabilities $p_{hj}(s, t|Z) = \mathbb{P}(X(t) = j | X(s) = h, Z)$ $s \leq t$ and $h \leq j$ for a given covariate vector Z , are expressed in the following way:

$$\widehat{p}_{11}(s, t|Z) = \prod_{s < u \leq t} \left(1 - \sum_{j=2}^3 d\widehat{A}_{1j}(u|Z) \right), \quad \widehat{p}_{22}(s, t|Z) = \prod_{s < u \leq t} \left(1 - d\widehat{A}_{23}(u|Z) \right)$$

$$\widehat{p}_{12}(s, t|Z) = \sum_{u \leq t} \widehat{p}_{11}(s, u - |Z) d\widehat{A}_{12}(u|Z) \widehat{p}_{22}(u +, t|Z).$$

where $\widehat{A}_{hj}(t|Z) = \widehat{A}_{hj0}(t) \exp\left(\widehat{\beta}_{hj}^T Z\right)$ is the estimates of the cumulative intensity function with $\widehat{A}_{hj0}(\cdot)$ the Breslow estimator for $A_{hj0}(t) = \int_0^t \alpha_{hj0}(u) du$. That is, estimates from the Cox models are simply plugged into (2)–(4).

One alternative approach is to use a semi-Markov model in which the future of the process does not depend on the current time but rather on the duration in the current state. Semi-Markov models¹² are also called “clock reset” models, because each time the patient enters a new state time is reset to 0. In this way, Cox semi-Markov models (CSMM) can be easily fitted (for the illness-death model, the only difference between CMM and CSMM is on transition $2 \rightarrow 3$), and so no further details are given here. General semi-Markov models may be analysed using Cox type models where duration effects are modelled via time-dependent covariates as exemplified by, for example Andersen *et al.*¹² and Andersen and Keiding.⁸

2.3.2 Introducing flexible covariate effects—For both the Cox proportional hazards model (9) and the additive hazards model (10) the effect of prognostic factors is assumed to have a linear (or log-linear) functional form. However, if this assumption is violated it may lead to erroneous statistical conclusions: bias and decreased power of tests for statistical significance.²⁶ In the Cox framework, the incorrect functional form for a covariate can also lead to a diagnosis of non-proportional hazards.²⁷ The need to relax this functional form has led to many further developments in survival analysis. The implementation of these methods in the framework of the MSMs can easily be considered for Cox-like models presented in Section 2.3.1 of this article. Software for fitting such methods is available as part of S-plus and R statistical languages. The issue of how to implement non-linear covariate effects for other multi-state approaches is not straightforward.

Within the context of the Cox proportional hazards model, several approaches to the problem of testing linearity have been proposed. A general proportional hazards model with an arbitrary covariate effect is of the form

$$\alpha(t; Z) = \alpha_0(t) \exp(f(Z))$$

where $f(Z)$ is assumed to be a smooth function of Z . Tibshirani and Hastie²⁸ considered non-parametric function estimation to this problem. Although this model does not confine the hazard to be log linear in Z , it is difficult to interpret the influence of any single covariate on survival. Furthermore, when Z has many components this approach is subjected to the “curse of dimensionality” and thus is not recommended. This problem can be overcome reducing the dimensionality by considering an additive regression model, expressing the log hazard as an additive function of each covariate:

$$\alpha(t; Z) = \alpha_0(t) \exp\left(\sum_{j=1}^p f_j(Z_j)\right)$$

where $f_j(\cdot)$, $j = 1, \dots, p$ are unspecified smooth covariate functions. These models have been studied by several authors using various non-parametric techniques. Spline-based smoothing methods were considered, for instance, by O'Sullivan,²⁹ Hastie and Tibshirani,³⁰ Gray³¹ and Kooperberg *et al.*³² is achieved by imposing a penalty for curve roughness. These models have gained recent popularity due to Eilers and Marx³³ when they introduced penalties to the B-splines. Huang *et al.*³⁴ used functional ANOVA decompositions to study a general class of models that includes the additive proportional hazards model. Recently, Huang and Liu³⁵ considered a proportional hazards model of the form

$$\alpha(t; Z) = \alpha_0(t) \exp(\psi(\beta^T Z))$$

where $\psi(\cdot)$, is an unknown smooth function. They first approximate the unspecified link function by a polynomial spline and then employ the maximum partial likelihood estimation.

These methods can be used to test for the presence of non-linear effects and to identify a “correct” functional form.

3 Existing software

While the Cox model with time-dependent covariates (TDCM) can be fitted through all the major statistical packages, MSMs need specialised software, most of which are written in FORTRAN, R or SAS. In the nineties, Marshall *et al.*³⁶ developed a FORTRAN program called *MARKOV*, for fitting general k -state Markov models (with $k - 1$ transient states and where the k th state can be optionally chosen as an absorbing state) with constant transition intensities and covariates. Later, Alioum and Commenges³⁷ presented a new computer program, called *MKVPCI*, which extends *MARKOV* by allowing piecewise-constant intensities with different values in at most three time intervals. This method needs prior specification of the cutpoints. The output of this software includes estimates of the baseline transition hazards, regression parameter estimates, transition intensities in time intervals together with their estimated standard errors, and the results of the multivariate hypothesis tests. However, it does not supply graphical output.

Jackson³⁸ developed the R package *msm*, implementing several functions for fitting continuous-time Markov and hidden Markov multi-state models (a model in which the stages are observed with misclassification) to longitudinal data. Covariates can be fitted to both the transition rates and misclassification probabilities. A variety of observation schemes are supported, including processes observed at arbitrary times, completely-observed processes, and censored states. Any pattern of transitions between states can be specified. The *msm* package provides several numerical outputs such as estimated coefficients (with confidence intervals), time spent in each state, the estimate of the ratio of two estimated transition intensities etc. Graphical output includes the survival plot and plot of the expected probability of survival against time (from each transient state). Tables of observed and expected prevalences (which can be used as a rough indication of the goodness of fit) can also be obtained.

Methods introduced by Therneau and Grambsch²⁷ can also be used to model multi-state survival data. These methods are based on Cox's regression model and can be fitted using standard software (R, S-plus, SAS, etc.). The library *survival* available as part of S-plus and R statistical packages can be used to implement these methods. The key here is the creation of an appropriate data set representing each individual by several observations. This approach can be done for Markov and semi-Markov models and can deal with any kind of process though it becomes complicated with the increase of the number of states. More details about this

procedure will be given later in our applications. For recurrent events, these methods can also be fitted using an SAS macro, called *PTRANSIT*, developed by Paes and Lima³⁹ for Markov processes. This macro can also be used for estimating transition probabilities.

Hui-Min *et al.*⁴⁰ developed a SAS macro for estimating the transition parameters in non-homogeneous (Weibull distributions, log-logistic, etc.) k -state progressive Markov model with several covariates.

Rosthøj *et al.*⁴¹ developed two SAS macros for estimation of the transition probabilities for a CMM for competing risks survival data. A competing risks model with k causes of failures is considered.

Recently, Wangler *et al.*⁴² developed a R library called *changeLOS* that implements the Aalen–Johansen estimator for general multi-state models with non-parametric hazards. The main goal of *changeLOS* is to compute the change in LOS (length of hospital stay), frequently used to assess the impact and the costs of hospital-acquired complication. So far, no covariates are included in this library.

Most of the existing software for MSMs presents, however, some difficulties and limitations in practice. Most of the available software assumes the process to be Markovian and/or time-homogeneous, and some of them are available as part of statistical packages which are not freely downloadable. Furthermore, possible comparisons between different multi-state models are rather difficult because each of the current programs requests its own data structure.

We, therefore, developed a user-friendly R library, that we called *tdc.msm*, for the analysis of multi-state survival data. Specifically, the new software may be used to fit the TDCM but also the reviewed MSMs (THMM, NHM, CMM and CSMM). In the models considered, patients pass from the initial state through one of a set of mutually exclusive states to an absorbing state. Advantages of this software include the same data input for fitting the different models while providing the corresponding numerical and graphical outputs obtained: parameter estimates with standard errors for the covariates; transition rates; survival estimates; transition probabilities estimates; and flexible p-spline hazard ratio estimates for continuous covariates.³³ Moreover, users are able to include any number of covariates on transition intensities.

Our library allows users to choose between Markov (THMM, NHM, CMM) and semi-Markov models (CSMM), and between time-homogeneous (THMM) and non-homogeneous models. A homogeneous Markov model can be fitted using the *msm* R package, a non-homogeneous piecewise model using *MKVPCI*, or a CMM using *PTRANSIT*, but our software puts them into the same library. In this way, users may easily analyse the results offered by the various models in order to compare them and make decisions accordingly. The *tdc.msm* program can be downloaded free of charge from <http://www.mct.uminho.pt/lmachado/Rlibrary>. Technical description of this program is provided in the independent article Meira-Machado *et al.*⁴³

Examples of analysis using this software will be shown in the following section using two data sets: the well-known Stanford heart transplant study data, and the Galician breast cancer data set.

4 Examples

4.1 Stanford heart transplantation study

For illustration, we apply several of the proposed methods of Section 2 to data from the Stanford heart transplant study. This study began in October 1967. The available data in Crowley and Hu's article⁴⁴ covers the period until 1 April 1974. Some patients died before an appropriate heart was found. Of the 103 patients, 69 received a heart transplant. The number of deaths was

75; the remaining 28 patients contributed with censored survival times. For each individual, an indicator of its final vital status (censored or not), the survival times (time to transplant, time to death) from the entry of the patient in the study (in days), and a vector of covariates including age at acceptance (age), year of acceptance (year), previous surgery (surgery: coded as 1 = yes; 0 = no), and transplant (coded as 1 = yes; 0 = no) were recorded. The covariate transplant is the only time-dependent covariate, while the other covariates included are fixed.

4.1.1 Time-dependent Cox regression model—One issue in the framework of the Cox model is the choice of time scale.⁴⁵⁻⁴⁷ For the Stanford data there are two choices of time scale for a Cox regression model: time-on-study and age. The standard approach for survival analysis is to define the survival time as the elapsed time from entry into the study until death, and to adjust by age using regression procedures. Using the Stanford heart transplantation data, we illustrate the features of the Cox model using the two time scales.

The analysis of the Cox model with time-dependent covariates (TDCM) can be obtained using almost all the existing statistical packages. To accommodate time-dependent effects, the S-PLUS/R statistical packages use a counting process data-structure introduced by Andersen and Gill.⁴⁸ In this data-structure, an individual's survival data is expressed by three variables: *start*, *stop* and *event*. For the Stanford study, the time-dependent covariate “transplant” represents a treatment intervention. Individuals without change in the time-dependent covariate are represented by only one line of data, whereas patients with a change in the time-dependent covariate must be represented by two lines. For these patients, the first line represents the time period until the transplant; the second line represents the time period that passes from the transplant to the end of the follow-up. The remaining (time-fixed) covariates are the same for the two lines. For each row, variables *start* and *stop* mark the time interval (*start*, *stop*) for the data, while *event* is an indicator variable taking on value 1 if there was a death at time *stop*, and 0 otherwise.

The most common type of time-dependent covariates are repeated measurements on a subject or a change in the subject's treatment. Both of these situations are easily handled by the counting process formulation. As an example consider the information available from four patients (from the Stanford study) with identification 25, 26, 27 and 28. For the first two patients the time from enrolment to censoring is 1800 and 1401 days, respectively, and the first patient had a heart transplant 25 days after enrollment. The time from enrolment to death for the third and fourth patients are 263 and 72 days, respectively, and the last patient received a new heart at day 71. The data for these four patients is represented as follows

id	start	stop	event	transplant	age	year	surgery
25	0	25	0	0	33.2238	1.57426	0
25	25	1800	0	1	33.2238	1.57426	0
26	0	1401	0	0	30.5353	1.58248	0
27	0	263	1	0	8.7858	1.59069	0
28	0	71	0	0	54.0233	1.68378	0
28	71	72	1	1	54.0233	1.68378	0

This approach for representing standard survival data can be easily extended to more complex situations. We shall return to this basic representation later.

We first construct various TDCMs all of them including the effect of transplant, among other covariates, and using time-on-study as the time-scale. Results for such a model can be obtained using the counting process data-structure introduced above. This can be done using Therneau's *survival* library as follow:

```
> coxph(Surv(start, stop, event) ~ transplant + age + year + surgery, data = heart)
```

Or using our library *tdc.msm* (with the same input data file) through the following input command:

```
> tdc.msm(heart, formula = c(6,7,8), models = "TDCM")
```

where the formula indicates the columns occupied by the time-fixed covariates to be included in the model. Numerical results obtained (not shown) are in agreement with those found by Crowley and Hu⁴⁴ and Kalbfleisch and Prentice.⁴⁹ In all the fitted models, the influence of age at acceptance on hazard is positive, while effects of year and surgery are both negative. When analysing models showing smaller AIC (Akaike's Information Criterion) we see that the effect of the transplant leads to a small reduction in risk, but without reaching statistical significance (Hazard Ratio, HR: 0.990; 95% confidence interval, 95%CI: 0.535–1.831). Age (HR: 1.027; 95%CI: 1.001–1.055) and year of acceptance (HR: 0.864; 95%CI: 0.753–0.992) are both important factors, while surgery has no significant effect ($P > 0.05$).

In addition, we construct TDCMs using age as the time-scale (rather than time-on-study). Note that in such models the risk set becomes everyone who was at risk at a certain age rather than at a certain event time. Surprisingly, this approach yields quite different estimates for the transplantation covariate. Now the effect of the transplant leads to an important reduction in risk, being the most important predictor of survival (HR: 0.296; 95%CI: 0.180–0.487), whereas year of acceptance and surgery have no significant effect. Results obtained from the two approaches illustrate that a careful choice of the time-scale is very important. The most likely explanation for the differences between the two Cox models (with different time-scale) seems to result from a non-linear effect of age in the TDCM with time-on-study as the time-scale (results not shown). This effect is closer to being quadratic, as previously suggested by Crowley and Hu.⁴⁴ When using time-on-study as the time scale in the Cox model, the effect of age is modelled parametrically by including it as a covariate. This assumes that the risk increases exponentially with age. Since this assumption is not valid, this will produce correlated residuals which may also be related to the time-dependent covariate. A more formal analysis to examine the extent to which the effect of transplant was affected by the choice of the time scale is required. Further details on this issue are out of the scope of the present work and they will be reported elsewhere.

As in our applications, general survival analysis problems often involve categorical time-dependent covariates. These covariates can be re-expressed as a MSM with states based on the values of the covariates. For the Stanford data set, "transplantation" (coded as 1 = yes; 0 = no) is the only time-dependent covariate and makes it feasible to use a three-state illness-death model to study the clinical progression of the disease. The influence of these intermediate events on survival is often important and can be handled using TDCM. In fact, such a TDCM can be the basis for a very flexible MSM like model depicted in Figure 2 where the transitions $1 \rightarrow 3$ and $2 \rightarrow 3$ are assumed proportional and where transition $1 \rightarrow 2$ is not modelled. This (less ambitious) partial MSM is obtained by adding interactions with the time-dependent covariate. An interaction between the time-dependent covariate (transplant) and the time-fixed covariate, Z , will model the situation where Z has different effects before and after the transplant. However, the inclusion of information on the intermediate event in a general multi-state model may give rise to better estimators for those effects. These models provide important information by highlighting covariates affecting both mortality and recurrence. These modelling approaches will now be considered in detail.

4.1.2 Multi-state models: the Markov assumption—In the context of multi-state modelling, we may consider the covariate “transplant” as an associated state of risk, and then use the progressive illness-death model with states “own heart” (having his/her own heart), “new heart” (or transplantation) and “dead”. Typically, a patient enters the study in the “own heart” state; after the transplant, he moves to the transplanted population, that is, to the “new heart” state. With this multi-state formulation of the Stanford data, and using time-on-study as the time-scale (in the following, this will be the case), main goals of this study include: (a) to assess whether or not a beneficial effect of heart transplant on survival exists. This will be carried out by comparing the transition intensities $\alpha_{13}(t)$ and $\alpha_{23}(t)$; and also (b) to explore the potential fixed covariate effects on each of the transitions.

Before using MSMs, we have to evaluate whether the Markov assumption is tenable. The Markov assumption is that future evolution only depends on the current state at time t ; in other words, the history of the process is summarised by the state occupied at time t . The Markov assumption may be checked, among others, by including covariates depending on the history.¹² For the illness-death model (the Markov assumption is only relevant for transition $2 \rightarrow 3$), we can examine whether the time spent in the healthy (“own heart”) state (i.e. the past) is important on the transition from the disease (“new heart”) state to death (i.e. the future). For doing that, let Z = “time spent in state 1” (alternatively, we could use “time since entry in state 2”), and t the current time. Fitting a model $\alpha_{23}(t; Z) = \alpha_{23,0}(t) \exp\{\beta Z\}$, we now need to test the null hypothesis, $H_0: \beta=0$, against the general alternative, $H_1: \beta \neq 0$. This would assess the assumption that the transition rate from the disease state into death is unaffected by the time spent in the previous state.

Following this procedure we verify that the effect of time spent in state 1 is not significant ($P > 0.05$) proving no evidence against the Markov model for the Stanford data.

4.1.3 Cox–Markov Model—CMMs can be fitted through most of the statistical packages as long as we use a counting process notation, representing each patient by several observations.²⁷ For the Stanford heart transplant study, individuals without transplant contribute with two lines of data (one for each of the transitions leaving state 1) whereas individuals with a transplant contribute with three lines of data (one for each transition). The counting process data-structure has now one more variable representing the transition.

For example, consider the same four patients from the Stanford heart transplantation study previously used. The same data is now represented using a long format like this:

id	start	stop	event	transplant	age	year	surgery	transition
25	0	25	0	0	33.2238	1.57426	0	1
25	0	25	1	0	33.2238	1.57426	0	2
25	25	1800	0	1	33.2238	1.57426	0	3
26	0	1401	0	0	30.5353	1.58248	0	1
26	0	1401	0	0	30.5353	1.58248	0	2
27	0	263	1	0	8.7858	1.59069	0	1
27	0	263	0	0	8.7858	1.59069	0	2
28	0	71	0	0	54.0233	1.68378	0	1
28	0	71	1	0	54.0233	1.68378	0	2
28	71	72	1	1	54.0233	1.68378	0	3

In this data-structure, *transition* = 1 denotes the mortality transition without transplantation, *transition* = 2 denotes the transplantation transition and *transition* = 3 the mortality transition after the transplant. The variable *event* denotes whether the event time (of interest) is observed or censored. The events of interest are death without transplant, transplant, and death after transplant, respectively.

The CMM allows us to observe how the covariate effect behaves when transition intensities are modelled separately. Again, we can choose between the “*survival*” library and our library “*tdc.msm*”. While the *survival* library requires some extra effort with respect to data preparation (the long format data-structure shown above), the *tdc.msm* library use the same data set previously used for Cox analyses. We can use Terry Therneau's library as follow (using the new data set with long format, say heart2):

```
> coxph(Surv(start,stop,event)~ age+year+surgery, data=heart2, subset=c
(transition==1))

> coxph(Surv(start,stop,event)~ age+year+surgery, data=heart2, subset=c
(transition==2))

> coxph(Surv(start,stop,event) ~ age+year+surgery, data=heart2, subset=c
(transition==3))
```

Using *tdc.msm* the following input command provides the (same) results for all the transitions

```
> tdc.msm(heart, formula=c(6,7,8), models= “CMM”)
```

As a consequence all the covariates are in the model. Table 1 presents the corresponding output for the CMM, which performs also a test for checking the Markov assumption (as discussed before). Smoothed (P-spline) log-hazard ratio curves can be obtained for continuous covariates using an extra argument *graphcov=T*. The results obtained from fitting this model show us that year of acceptance, which revealed a strong effect on survival in the Cox model, under the CMM only obtains a significant effect on $\alpha_{13}(t)$. Age turns out to be the best predictor for the mortality transition $\alpha_{23}(t)$ of transplanted patients, and also for transition $\alpha_{12}(t)$, corresponding to receive a new heart. The effect of age on the mortality intensity in patients without transplant is not significant. No significant effect of a previous surgery is found.

4.1.4 Time Homogeneous Markov Model—Like the preceding studied models, the THMM offers a detailed description of the survival process, making use of all the available information to estimate the effect of prognostic factors and intensity rates. By applying this modelling approach, we refit the Stanford data including the potential effects of age, year and surgery in all transitions. Results obtained from the fitted model (Table 2) are in good agreement with those obtained through the CMM. An exception is the surgery effect, which was not a significant factor in any of the previously studied models, and now turns out to be significant for transplanted patients (HR: 0.266; 95% CI: 0.112–0.631). Results indicate that age is the only covariate showing a significant linear effect in all transitions. This occurs even for the mortality intensity in patients without a transplant, something that did not occur when using the Cox Markov modelling approach. We also observe that the acceptance time in the study is a significant predictor, though only for the mortality intensity in patients without transplant (HR: 0.643; 95% CI: 0.450–0.920). These results were obtained using the following input command:

```
> tdc.msm(heart, formula=c(6,7,8), models= “HMM”)
```

In addition to the results presented in Table 2, the function shown above provide mean sojourn times (in each transient state), prevalence tables (if an extra argument *prevalence=T* is used) and graphical output (survival estimates and transition probabilities estimates if *surv.plot=T* and *plot.trans=T* is used).

Also in the THMM it is very simple to carry out a formal test to compare effects of the same prognostic factor (transition rates also) on the transition intensities. For example, we can see from Table 2 that the fitted THMM leads to very similar effects of age in the three transitions. To assume that these effects are equal, we use of the Wald test statistic, yielding a value of 0.614, revealing non-significant differences between them. Such a test is a part of the output for the above input command.

In view of the results obtained, we consider a simplified homogeneous Markov model, including the (same) potential effect of age in all transitions, the effect of year only in the mortality transition of patients without transplant, and the effect of surgery in the mortality transition of transplanted patients. In the process, likelihood ratio tests are used to test if the regression parameters are statistically different from zero. This statistic has an approximately χ^2_1 distribution under $H_0 : \beta_{hj} = 0$. Estimated parameters for the simplified model (results not shown) do not differ substantially from those obtained by the initial model (shown in Table 2).

Further, we use Wald's test to study whether or not a relation between transplant and survival exists. Formally, the hypothesis of no relation is given by $H_0 : \alpha_{13} = \alpha_{23}$, and then Wald's test reduces to $W = (\hat{\alpha}_{13} - \hat{\alpha}_{23})^2 / v_{11}$, being $v_{11} = \text{var}(\hat{\alpha}_{13} - \hat{\alpha}_{23})$. With our data, under the null hypothesis the W statistic (which follows a χ^2_1) yields a value of 3.121, suggesting that the transplant leads to a reduction in risk but without reaching statistical significance ($P = 0.077$). Note again that likelihood ratio tests can also be used for constructing a test of H_0 .

The results obtained by comparing the observed and the predicted number of patients occupying each state at particular times (not reported here) reveal that, for lower survival times, mortality is underestimated from the fitted homogeneous Markov model. In many cases these discrepancies can be explained by the failure of the Markov assumption (which is not the case here). Another possibility is that the transition rates vary with time, so that the model is non-homogeneous. We believe this is the case for the Stanford data: it is seen that most of the transitions from state “own heart” to state “new heart” (approximately 73% from the total) occur up to 51 days of survival. Taken as a whole, these results suggest that a homogeneous model may be inappropriate. To assess the assumption of time homogeneity, Kay¹⁰ suggests the use of a piecewise constant model. Again, likelihood ratio tests can be used to compare the piecewise constant model with the homogeneous model. Such a model will now be constructed.

4.1.5 Piecewise constant intensities model—In this section we build a piecewise constant intensities model (described in Section 2.1.2) with one cutpoint, specified from the Stanford data covariates that showed a significant effect when fitting the homogeneous model. After examining the likelihood for several cutpoints θ , a value of $\theta = 80$ days was selected, and two intervals (time ≤ 80 days, time > 80 days) were then considered. The choice of appropriate cutpoints is a very important (but rather difficult) issue in NHMs. For example, this choice should make sure that an appropriate (sufficient) number of observations fall in all intervals.

For obtaining the numerical results for the NHM with a cut-off point of 80 days we can use the following input command

```
> tdc.msm(heart, formula=c(6,7,8), models="NHM", cut=80)
```

Such a model reveals that the covariate surgery is not important in any transition. A new model is fitted excluding this covariate

```
> tdc.msm(heart, formula=c(6,7), models="NHM", cut=80)
```


From this new model we verify that in both intervals the differences between the estimates for both mortality intensities are not statistically significant. When examining the fixed covariate effects, we see that, for time ≤ 80 days, year of acceptance is a significant predictor in transitions from state 1 to state 3 (HR: 0.673; 95%CI: 0.465–0.976). For the second interval (time > 80 days), however, the only significant covariate is age at acceptance for the transition from state 1 to state 2 (HR: 1.209; 95%CI: 1.027–1.425).

4.1.6 Diagnostics for model assessment—The Markov property and the homogeneity assumption are strong assumptions which may lead to biased estimates if violated. To our knowledge, the *tdc.msm* library is the only available software which provides a test for the Markov assumption. To assess the homogeneity assumption likelihood ratio tests can be used to compare the piecewise constant model with the homogeneous model. None of the available software's performs such a test though it can be easily performed by fitting the two models and comparing their likelihoods.

Also important is the issue of goodness-of-fit of a model. One may attempt to address goodness-of-fit by comparing the observed and predicted number of individuals occupying each state at a particular time.⁵⁰ The method proposed can be used for irregular sampling times when no explanatory variables are included in the model. When exact time transitions are observed the fit of the expected number of individuals in each state can be assessed directly at those times. However, in medical applications, the data is incomplete as the process is observed only at fixed times and transition times are interval censored. This can lead to an additional problem since the precise state each subject occupies at the assigned times will not be known. The method use approximated counts calculated assuming that subjects would remain in the state they were observed at in the previous observation. An indication of where the data deviate from the model is obtained by comparing the observed number of individuals occupying state h , $O_h(t)$, with the expected number of patients, $E_h(t)$, under the fitted model, for a particular time t through $M_h(t) = (O_h(t) - E_h(t))^2/E_h(t)$. However, there is no formal way of assessing if the deviances are statistically significant. Such a rough indication of the goodness-of-fit of the model (through prevalence counts) can be obtained for the THMM using either *tdc.msm* (putting an extra argument, prevalence=T, in the last input command) or using Jackson's R library *msm*.

We note that Aguirre-Hernandez and Farewell⁵¹ developed a Pearson-type test for stationary and continuous time Markov models, which can be used when the transition rates depend on several explanatory variables. Since this method is based on partitioning the data set on the basis of interval observations, their methods is not applicable when the process includes an absorbing state for which the time of entry is known exactly. A recent paper by Titman and Sharples⁵² presents a modification to the Pearson-type test to allow for these cases.

We have illustrated several of the methods in Section 2.1 using the Stanford data. Previous analysis of this data set in the medical literature focused on the effect of the covariates. Among others, Turnbull, Brown and Hu,⁵³ Mantel and Byar⁵⁴ and Crowley and Hu⁴⁴ studied the Stanford heart data, reaching a small negative influence of transplantation on hazard, but without statistical significance. We have obtained similar results by using multi-state models. Multi-state modelling has also shown that year of acceptance (i.e. fixed covariate) affects each of the transition intensities in a different way.

4.2 Breast cancer data

In breast cancer research, patients who experienced a recurrence (local-regional or metastases) have a significantly higher mortality risk. In such cases one main interest is to identify which prognostic factors are predisposing to recurrence. Furthermore, it is important to investigate how these factors behave in the disease process. Recurrence affects the patient's outcome and

can be included as a transient state in a progressive three-state model with states “alive and disease-free”, “alive with recurrence” (local-regional or metastases) and “death”.

In the period between April 1991 and December 2003, 585 patients with breast cancer were treated in Galicia (Spain). From the total of the patients, 172 relapsed (recurrence) and among these 114 died. Eleven patients died without relapse. These patients are included in the TDCM but in the MSM they are treated as censored on the recurrence transition and they are not considered on the mortality transition from the “alive with recurrence” state. The rest of the patients remained alive and disease-free up to the end of the follow-up. Patients were monitored on a regular basis and information was recorded for each subject, including an indicator of her final status, the survival times from the entry in study, for recurrence and death, and a vector of covariates including age, tumour size (cm), tumour perimeter (mm), and three covariates (radiotherapy, chemotherapy and hormonotherapy) representing different treatments (coded as 1 = yes and 0 = no).

Results obtained using Cox time-dependent regression model confirm that patients with recurrence have significantly poorer prognosis (HR: 18.777; 95%CI: 8.693–40.56). Furthermore, tumor size (HR: 1.082; 95%CI: 1.034–1.13), age (HR: 1.026; 95%CI: 1.009–1.04), and perimeter (HR: 1.031; 95%CI: 1.019–1.04) reveal to be important prognostic factors of survival. None of the treatments obtain a statistical significant effect on survival.

Again, the Markov assumption can be checked by including covariates depending on the history. For this application we verify a strong (negative) effect of time since the recurrence on the mortality transition. In such cases, there is a duration effect of $\alpha_{23}(\cdot)$, and so, semi-Markov models are often considered. Here, we consider Cox semi-Markov models (CSMM) to investigate the covariate effects on survival. These models can be fitted using the counting process data-structure as shown above for the CMMs as follows

```
> coxph(Surv(stop-start, event) ~ perimet+size+age+radio+chemo+hormone,
data=breast, subset=(transition==1))

> coxph(Surv(stop-start, event) ~ perimet+size+age+radio+chemo+hormone,
data=breast, subset=(transition==2))
```

The same results are obtained using library *tdc.msm* with

```
> tdc.msm(breast, formula=c(6,7,8,9,10,11), models="CSMM")
```

Table 3 summarises results after fitting the CSMM. Results show that some of the significant covariates in the simple Cox model turn out to affect the transition intensities differently, though the effects on the mortality transition ($2 \rightarrow 3$) are in good agreement with those obtained using the simple Cox model. For example, the covariate age, which revealed to be an important prognostic factor in the simple Cox model, only obtains a significant effect on $\alpha_{23}(\cdot)$. Furthermore, the covariate effect on transition $\alpha(\cdot)$ is opposite to that on mortality transition. Results also indicate a significant effect of chemotherapy on recurrence transition.

For continuous covariates it is possible to test whether the covariate effect is linear or not. Fitting a Cox model with penalised splines,³³ and using the number of degrees of freedom which minimises AIC, reveals a non-linear covariate effect for perimeter on the recurrence transition ($P < 0.05$). There is no evidence of non-linear effect on the mortality transition. Results for such a test as well as the graphical output can be obtained including an extra argument in the last input command,

```
> tdc.msm(breast, formula=c(6,7,8,9,10,11), models="CSMM", graphcov = 1)
```

Figure 4 displays the penalised spline fit for perimeter with the 95% pointwise confidence limits. This plot suggests that the functional form for perimeter is rather constant until around 60 and then rapidly increase until 95, remaining nearly constant afterwards.

With this application we are also particularly interested in illustrating differences between the estimated transition probabilities from Aalen–Johansen estimator (Markovian) and from “Markov-free” estimator of Meira-Machado *et al.*²⁵ In Figure 5 we present, as an example, estimated transition probabilities for $p_{hj}(2, t)$ and $p_{hj}(6, t)$, $h = 1, 2, j = 1, 2, h \leq j$, for the breast cancer data, showing that a choice between those two approaches makes a big difference. From these figures we can see more clearly the effect of the intermediate event (recurrence) in the patient survival prognosis, showing a much poorer survival prognosis for those individuals in state 2.

As it can be seen from Figure 5, the “Markov-free” estimator has fewer jump points but bigger steps. The number of jump points and the size of the steps are related to censoring and to the sample size. Also, we observe important differences between the estimated survival curves for individuals who have had recurrence, differences which became clearer with the progression of time, showing that the estimated prognosis using the “Markov-free” method is poorer than based on the Aalen–Johansen approach.

Note that since the process for the breast cancer data does not fulfill the Markov assumption (as shown in Table 3) the use of Markov-free estimators is preferable to Aalen–Johansen estimators.

5 Discussion

In this article, we have discussed the use of multi-state models in the analysis of survival data. We have illustrated the application of these methods using two data sets: the Stanford heart transplant study and a Galician data set on breast cancer, and we have provided a review of software including some flexible methods to cope with non-linear covariate effects and with non-Markov data structures. The potential advantages of these methods were illustrated through the referred data sets.

When analysing Stanford heart data the Markov assumption was tested showing that the transition rates in states are not affected by the previous sojourn time. Several classical Markov models to deal with such data are discussed, focussing on the estimation of the covariate effects. When comparing these multi-state approaches with the well known Cox model with time-dependent covariates we have illustrated that the multi-state modelling yielded new insights while confirming some of the results obtained with the Cox model.

We have also used a new data set on breast cancer to compare Cox regression analysis and multi-state approaches. Since for this application the Markov assumption is violated, methods not relying on the Markov property were used for the estimation of covariate effects (Cox semi-Markov model), as well as for the estimation of transition probabilities (non-Markov). Furthermore, we have verified that, on the recurrence transition, some covariate effects are best represented by smooth, non-linear functions.

In conclusion, multi-state modelling offers a flexible tool for the study of covariate effects on the various transition rates. These models may bring out important biological insights which may be ignored when using an ordinary regression model. In practice, MSMs can be used to confirm and thoroughly examine conclusions obtained by applying simpler survival models. Therefore, we should not see the MSMs as merely an alternative to the TDCM but rather as supplements that offer additional information.

Acknowledgements

The authors acknowledge financial support by Spanish Ministry of Education & Science grants MTM2005-01274 and MTM2005-00818 (European FEDER support included), and by the Xunta de Galicia grants PGIDIT06PXIC208043PN, PGIDIT06PXIC300117PN and PGIDIT07PXIB300191PR. Per Kragh Andersen was supported by a grant (R01 CA54706-12) from the National Cancer Institute and by the Danish National Research Council, grant 272-08-0472: Point processes and their statistical inference. We are also grateful to Professor Puente Dominguez (in memoriam) and to Dr Elisabeth Chávez for the use of the breast cancer data. Thanks to two anonymous referees for comments and suggestions which have improved the presentation of the article.

References

- Andersen PK, Abildstrom SZ, Rosthøj S. Competing risks as a multi-state model. *Statistical Methods in Medical Research* 2002;11:203–15. [PubMed: 12040697]
- Hougaard P. Analysis of multivariate survival data. Springer; 2000.
- Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data. Springer; 2003.
- Commenges D. Inference for multi-state models from interval-censored data. *Statistical Methods in Medical Research* 2002;11:167–82. [PubMed: 12040695]
- Andersen PK, Borgan O, Gill RD, Keiding N. Statistical models based on counting processes. Springer; 1993.
- Commenges D. Multi-state models in epidemiology. *Lifetime Data Analysis* 1999;5:315–27. [PubMed: 10650740]
- Hougaard P. Multi-state models: a review. *Lifetime Data Analysis* 1999;5:239–64. [PubMed: 10518372]
- Andersen PK, Keiding N. Multi-state; models for event history analysis. *Statistical Methods Medical Research* 2002;11:91–115.
- Chiang CL. Introduction to stochastic processes in biostatistics. Wiley; 1968.
- Kay R. A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics* 1986;42:855–65. [PubMed: 2434150]
- Cox DR, Miller HD. The theory of stochastic processes. Chapman & Hall; 1965.
- Andersen PK, Esbjerg S, Sorensen TIA. Multistate models for bleeding episodes and mortality in liver cirrhosis. *Statistics in Medicine* 2000;19:587–99. [PubMed: 10694738]
- Pérez-Ocón R, Ruiz-Castro JE, Gámez-Pérez ML. A piecewise Markov process for analysing survival from breast cancer in different risk groups. *Statistics in Medicine* 2001;20:109–22. [PubMed: 11135351]
- Saint-Pierre P, Combescure C, Daurès JP, Godard P. The analysis of asthma control under a Markov assumption with use of covariates. *Statistics in Medicine* 2003;22:3755–70. [PubMed: 14673936]
- Aalen O, Johansen S. An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* 1978;5:141–50.
- Frydman H. Nonparametric estimation of a Markov illness-death process from interval-censored observations, with application to diabetes survival data. *Biometrika* 1995;82:773–89.
- Joly P, Commenges D, Helmer C, Letenneur L. A penalized likelihood approach for an illness-death model with interval censored data: application to age-specific incidence of dementia. *Biostatistics* 2002;3:433–43. [PubMed: 12933608]
- Keiding N, Klein JP, Horowitz M. Multi-state models and outcome prediction in bone marrow transplantation. *Statistics in Medicine* 2001;20:1871–85. [PubMed: 11406847]
- Kalbfleisch JD, Lawless J. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* 1985;80:863–71.
- Pepe MS. Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association* 1991;86:770–78.
- Strauss D, Shavell R. An extended Kaplan-Meier estimator and its applications. *Statistics in Medicine* 1998;17:971–82. [PubMed: 9612885]

22. Aalen O, Borgan O, Fekjaer H. Covariate adjustment of event histories estimated from Markov chains: the additive approach. *Biometrics* 2001;57:993–1001. [PubMed: 11764270]
23. Datta S, Satten GA. Validity of the Aalen-Johansen estimators of stage occupation probabilities and Nelson-Aalen integrated transition hazards for non-Markov models. *Statistics and Probability Letters* 2001;55:403–11.
24. Glidden D. Robust inference for event probabilities with non-Markov event data. *Biometrics* 2002;58:361–68. [PubMed: 12071409]
25. Meira-Machado L, De Uña-Álvarez J, Cadarso-Suárez C. Nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis* 2006;12:325–44. [PubMed: 16917736]
26. Anderson GL, Fleming TR. Model misspecification in proportional hazards regression. *Biometrika* 1995;82:527–41.
27. Therneau, TM.; Grambsch, PM. Modeling survival data: extending the Cox model. Springer-Verlag; 2000.
28. Tibshirani R, Hastie T. Local likelihood estimation. *Journal of the American Statistical Association* 1987;82:559–67.
29. O'Sullivan F. Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal of Scientific and Statistical Computation* 1988;9:531–42.
30. Hastie TJ, Tibshirani RJ. Exploring the nature of covariate effects in the proportional hazards model. *Biometrics* 1990;46:1005–16. [PubMed: 1964808]
31. Gray RJ. Spline-based tests in survival analysis. *Biometrics* 1994;50:640–52. [PubMed: 7981391]
32. Kooperberg C, Stone CJ, Truong YK. Hazard regression. *Journal of the American Statistical Association* 1995;90:78–94.
33. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. *Statistical Science* 1996;11:89–121.
34. Huang JZ, Kooperberg C, Stone CJ, Truong YK. Functional ANOVA modelling for proportional hazards regression. *Annals of Statistics* 2000;28:960–99.
35. Huang JZ, Liu L. Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrics* 2006;62:793–802. [PubMed: 16984322]
36. Marshall G, Guo W, Jones RH. MARKOV: a computer program for multi-state Markov models with covariables. *Computer Methods and Programs in Biomedicine* 1985;47:147–56. [PubMed: 7587161]
37. Alioum A, Commenges D. MKVPCI: A computer program for Markov models with piecewise constant intensities and covariates. *Computer Methods and Programs in Biomedicine* 2001;64:109–19. [PubMed: 11137193]
38. Jackson CH, Sharples LD. Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine* 2002;21:113–28. [PubMed: 11782054]
39. Paes A, Lima A. A SAS macro for estimating transition probabilities in semiparametric models for recurrent events. *Computer Methods and Programs in Biomedicine* 2004;75:59–65. [PubMed: 15158048]
40. Hui-Min W, Ming-Fang Y, Hsiu-Hsi C. SAS macro program for non-homogeneous Markov process in modeling multi-state disease progression. *Computer Methods and Programs in Biomedicine* 2004;75:95–105. [PubMed: 15212852]
41. Rosthøj S, Andersen PK, Abildstrom SZ. SAS macros for estimation of the cumulative incidence functions based on a Cox regression model for competing risks survival data. *Computer Methods and Programs in Biomedicine* 2004;74:69–75. [PubMed: 14992828]
42. Wangler M, Beyersmann J, Schumacher M. Changelos: an R-package for change in length of hospital stay based on the Aalen-Johansen estimator. *R News* 2006;6:31–5.
43. Meira-Machado L, Cadarso-Suárez C, De Uña-Álvarez J. Tdc.msm: An R library for the analysis of multi-state survival data. *Computer Methods and Programs in Biomedicine* 2007;86:131–40. [PubMed: 17350136]
44. Crowley J, Hu M. Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association* 1977;72:27–36.

45. Korn EL, Graubard BI, Midthune D. Time-to-event analysis of longitudinal follow-up of a survey: choice of time scale. *American Journal of Epidemiology* 1997;145:72–80. [PubMed: 8982025]
46. Thiébaud A, Bénichou J. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics in Medicine* 2004;23:3803–20. [PubMed: 15580597]
47. Pencina MJ, Larson MG, D'Agostino RB. Choice of time scale and its effect on significance of predictors in longitudinal studies. *Statistics in Medicine* 2007;26:1343–59. [PubMed: 16955538]
48. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Annals of Statistics* 1982;10:1100–20.
49. Kalbfleisch, JD.; Prentice, RL. *The statistical analysis of failure time data*. John Wiley & Sons; 2002.
50. Gentleman RC, Lawless JF, Lindsey JC, Yan P. Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine* 1994;13:805–21. [PubMed: 7914028]
51. Aguirre-Hernández R, Farewell VT. A Pearson-type goodness-of-fit test for stationary and time-continuous Markov regression models. *Statistics in Medicine* 2002;21:1899–911. [PubMed: 12111896]
52. Titman AC, Sharples LD. A general goodness-of-fit test for Markov and hidden Markov models. *Statistics in Medicine*. 2007DOI 10.1002/sim.3033
53. Turnbull BW, Brown BW, Hu M. Survivorship analysis of heart transplant data. *Journal of the American Statistical Association* 1974;69:74–80.
54. Mantel N, Byar DP. Evaluation of response-time data involving transient states: an illustration using heart-transplant data. *Journal of the American Statistical Association* 1974;69:81–6.

Figure 1.
K-progressive model.

Figure 2.
Illness-death model.

Figure 3.
The bivariate model for bivariate parallel data.

Figure 4. Penalised spline fit for perimeter (mm) (with 95% pointwise confidence bands) for the recurrence transition. Breast cancer data.

Figure 5. Estimated transition probabilities for Aalen-Johansen estimator (solid line) and non-Markov model (dashed line). Breast cancer data.

Table 1
R output using tdc.msm for the Cox–Markov model. Stanford heart transplant data

COX-MARKOV MODEL. SEPARATED FITTED COX MODELS FOR EACH TRANSITION					
FROM STATE 1 TO STATE 2					
coef	exp(coef)	se(coef)	z	p-value	
age	0.03111	1.032	0.0140	2.2255	0.026
year	0.00075	1.001	0.0695	0.0108	0.990
surgery	0.04734	1.048	0.3152	0.1502	0.880
exp(coef)	exp(-coef)	lower.95	upper.95		
age	1.032	0.969	1.004	1.06	
year	1.001	0.999	0.873	1.15	
surgery	1.048	0.954	0.565	1.94	
Rsquare=0.054 (max possible=0.993)					
Likelihood ratio test=5.77 on 3 df, p=0.123					
Wald test=5 on 3 df, p=0.172					
Score (logrank) test=5.06 on 3 df, p=0.167					
-2*Log-likelihood: 509.5638					
FROM STATE 1 TO STATE 3					
coef	exp(coef)	se(coef)	z	p-value	
age	0.0198	1.020	0.0181	1.094	0.270
year	-0.2833	0.753	0.1110	-2.553	0.011
surgery	-0.2288	0.796	0.6361	-0.360	0.720
exp(coef)	exp(-coef)	lower.95	upper.95		
age	1.020	0.980	0.984	1.057	
year	0.753	1.328	0.606	0.936	
surgery	0.796	1.257	0.229	2.768	
Rsquare=0.08 (max possible=0.886)					
Likelihood ratio test=8.62 on 3 df, p=0.0347					
Wald test=8.19 on 3 df, p=0.0422					
Score (logrank) test=8.67 on 3 df, p=0.034					
-2*Log-likelihood: 214.9848					
FROM STATE 2 TO STATE 3					
Checking the Markov assumption:					

COX-MARKOV MODEL. SEPARATED FITTED COX MODELS FOR EACH TRANSITION
FROM STATE 1 TO STATE 2

	coef	exp(coef)	se(coef)	z	p-value
testing if the time spent in state 1 (start) is important on transition from state 2 to state 3					
coef	exp(coef)	se(coef)	z	p	
start	-0.0094	0.00534	-1.76	0.079	
Likelihood ratio test=4.5 on 1 df, p=0.0338					
coef	exp(coef)	se(coef)	z	p-value	
age	0.0496	1.051	0.0214	2.318	0.020
year	-0.0230	0.977	0.0969	-0.238	0.810
surgery	-0.8165	0.442	0.4549	-1.795	0.073
exp(coef)		exp(coef)	lower.95	upper.95	
age	1.051	0.952	1.008	1.10	
year	0.977	1.023	0.808	1.18	
surgery	0.442	2.263	0.181	1.08	
Rsquare=0.151 (max possible=0.987)					
Likelihood ratio test=11.3 on 3 df, p=0.0102					
Wald test=10.2 on 3 df, p=0.0167					
Score (logrank) test=10.7 on 3 df, p=0.0137					
-2*Log-likelihood: 290.1922					

Table 2
Sample of the R output using tdc.msm for the Time-Homogeneous Markov model

MULTI-STATE HOMOGENEOUS MARKOV MODEL				
Undergoing Transitions				
to				
from 1 2 3				
1 4 69 30				
2 0 24 45				
convergence				
-2*Log-likelihood: 1701.271				
Estimated coefficients				
age				
	Stage 1	Stage 2	Stage 3	
Stage 1	0	0.06898958	0.05024326	
Stage 2	0	0.00000000	0.07064008	
Stage 3	0	0.00000000	0.00000000	
year				
	Stage 1	Stage 2	Stage3	
Stage 1	0	0.02116241	-0.4412467	
Stage 2	0	0.00000000	0.1190527	
Stage 3	0	0.00000000	0.00000000	
surgery				
	Stage 1	Stage 2	Stage3	
Stage 1	0	0.006854728	0.4104957	
Stage 2	0	0.000000000	-1.3232050	
Stage 3	0	0.000000000	0.00000000	
baseline				
	Stage 1	Stage 2	Stage 3	
Stage 1	-0.02083436	0.016909053	0.003925302	
Stage 2	0.00000000	-0.002003426	0.002003426	
Stage 3	0.00000000	0.000000000	0.000000000	
age				
	HR	L95	U95	
Stage 1 - Stage 2	1.071425	1.042991	1.100634	
Stage 1 - Stage 3	1.051527	1.010531	1.094186	
Stage 2 - Stage 3	1.073195	1.028606	1.119717	
year				
	HR	L95	U95	
Stage 1 - Stage 2	1.021388	0.8883874	1.1743000	
Stage 1 - Stage 3	0.643234	0.4495542	0.9203562	
Stage 2 - Stage 3	1.126429	0.9514686	1.3335625	
surgery				
	HR	L95	U95	
Stage 1 - Stage 2	1.0068783	0.5386172	1.8822346	

MULTI-STATE HOMOGENEOUS MARKOV MODEL

Stage 1 - Stage 3	1.5075650	0.3489639	6.5128570
-------------------	-----------	-----------	-----------

Stage 2 - Stage 3	0.2662805	0.1123546	0.6310852
-------------------	-----------	-----------	-----------

Walds Test to check for differences between mortality intensities

Let H_0 be the hypothesis of non-differences between mortality intensities

H_0 produces a chisquare statistic of: 3.121163

with a p-value of: 0.07728162

Walds test to check for differences between covariate 1 in the transition intensities

Let H_0 be the hypothesis of non-differences between covariate 1 in the transition intensities

H_0 produces a chi-square statistic of: 0.6144772

with a p-value of: 0.7354751

Table 3
R output using tdc.msm for the Cox semi-Markov model. Breast cancer data

COX SEMI-MARKOV MODEL. SEPARATED FITTED COX MODELS FOR EACH TRANSITION					
FROM STATE 1 TO STATE 2					
	coef	exp(coef)	se(coef)	z	p-value
size	0.099248	1.104	0.02289	4.33665	1.4e-05
age	-0.000432	1.000	0.00706	-0.06113	9.5e-01
perimeter	0.055401	1.057	0.00495	11.19509	0.0e+00
radiotherapy	0.000811	1.001	0.20845	0.00389	1.0e+00
chemotherapy	-0.739912	0.477	0.24627	-3.00445	2.7e-03
hormonotherapy	0.206419	1.229	0.21321	0.96814	3.3e-01
	exp(-coef)		lower:95	upper:95	
size	1.104	0.906	1.056	1.155	
age	1.000	1.000	0.986	1.014	
perimeter	1.057	0.946	1.047	1.067	
radiotherapy	1.001	0.999	0.665	1.506	
chemotherapy	0.477	2.096	0.294	0.773	
hormonotherapy	1.229	0.813	0.809	1.867	
Rsquare=0.284 (max possible=0.966)					
Likelihood ratio test=135 on 6 df, p=0					
Wald test=155 on 6 df, p=0					
Score (logrank) test=188 on 6 df, p=0					
-2*Log-likelihood: 1233.925					
FROM STATE 2 TO STATE 3					
Checking the Markov assumption:					
testing if the time spent in state 1 (start) is important on transition from state 2 to state 3					
coef	exp(coef)	se(coef)	z	p	
start -0.494	0.61	0.0601	-8.23	2.2e-16	
Likelihood ratio test=95.2 on 1 df, p=0					
coef		exp(coef)	se(coef)	z	p-value
size	0.0776	1.081	0.02643	2.935	0.00330
age	0.0240	1.024	0.00898	2.669	0.00760
perimeter	0.0258	1.026	0.00711	3.621	0.00029

COX SEMI-MARKOV MODEL. SEPARATED FITTED COX MODELS FOR EACH TRANSITION
FROM STATE 1 TO STATE 2

	coef	exp(coef)	se(coef)	z	p-value
radiotherapy	-0.3549	0.701	0.24901	-1.425	0.15000
chemotherapy	0.1618	1.176	0.27826	0.581	0.56000
hormonotherapy	-0.0273	0.973	0.27210	-0.100	0.92000
	exp(coef)	exp(-coef)	lower.95	upper.95	
size	1.081	0.925	1.026	1.14	
age	1.024	0.976	1.006	1.04	
perimeter	1.026	0.975	1.012	1.04	
radiotherapy	0.701	1.426	0.430	1.14	
chemotherapy	1.176	0.851	0.681	2.03	
hormonotherapy	0.973	1.028	0.571	1.66	
Rsquare=0.227 (max possible=0.996)					
Likelihood ratio test=31.9 on 6 df, p=1.74e-05					
Wald test=32.7 on 6 df, p=1.19e-05					
Score (logrank) test=34.3 on 6 df, p=5.83e-06					
-2*Log-likelihood: 656.6577					