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

Luís Meira-Machado
Jacobo de Uña-Álvarez
Carmen Cardarso-Suárez
Per Kragh Andersen

Research Report 07/1
Department of Biostatistics
University of Copenhagen
**Multi-state models for the analysis of time to event data**

**Authors:** Luís Meira-Machado¹, Jacobo de Uña-Álvarez², Carmen Cadarso-Suárez³ and Per K. Andersen⁴

¹Department of Mathematics for Science and Technology, University of Minho, Portugal.

²Department of Statistics and Operations Research. University of Vigo, Spain.

³Department of Statistics and Operations Research, University of Santiago de Compostela, Spain.

⁴Department of Biostatistics, University of Copenhagen, Denmark.

**Name and address for correspondence:**

Luís Filipe Meira-Machado  
Department of Mathematics for Science and Technology  
University of Minho  
4810- Azurém, Guimarães. Portugal.  
Phone: +(351)253510400  
Fax: +(351)253510401  
E-mail: lmachado@mct.uminho.pt
Multi-state models for the analysis of time to event data

ABSTRACT. The experience of a patient in a survival study may be thought of as a process that involves two states, with one possible transition from a ‘live’ state to a ‘dead’ state. In some studies, however, the state representing those patients ‘alive’ may be partitioned into two or more intermediate (transient) states, each of which corresponding to a particular stage of the illness. In such studies where at any time point individuals occupy one of a set of discrete states, multi-state models can be used to model the movement of patients between the various states. These models may offer a better understanding of the process of the illness, leading to a better knowledge of the evolution of the disease over time. 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 several modelling approaches following the methodology of the multi-state models, and focusing on the estimation of several quantities such as the transition probabilities and survival probabilities. Differences between these approaches are discussed, focusing on possible advantages and disadvantages for each method. Software to analyse such models has been developed in the form of an R library. Different approaches and software are illustrated using data from the Stanford Heart Transplant Study and with data from a study on breast cancer conducted in Galicia, Spain.

Keywords: Proportional Hazards Model; Multi-state Model; Markov process; Time-dependent covariate.
1 Introduction

In longitudinal studies of disease, patients are observed and covariate information is collected at several occasions through a follow-up period. The analysis in such studies where individuals may experience several events is often performed using multi-state models. A multi-state model (MSM) is defined as a model for a continuous time stochastic process allowing individuals to move between 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 state structure identifies the states and also the transitions allowed between states. The complexity of a MSM greatly depends on the number of states defined and also on the transitions allowed between these states. Graphically these models are illustrated using boxes -to represent the states- and arrows going from one state to another.

The simplest form of MSM is the mortality model (with states “alive” and “dead” and only one transition allowed between them) for survival analysis. Splitting the “Alive” state from the simple mortality model for survival data into two transient states, we therefore obtain the simplest progressive three-state model. Both models are particular models of the k-progressive model, illustrated in Figure 1.
Another possible MSM to describe the disease progression is the *illness-death model* (also known as disability model). This model (Figure 2) is widely used in the medical literature and can be used to study the incidence of the disease as well as death. One important problem here may be to evaluate whether previously diseased subjects have the same risk of death as those who have been healthy all their lives.

![Figure 2: Illness-death model.](image)

The *competing risks model* (Andersen et al., 2002) 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* (Hougaard, 2000), depicted in Figure 3, is the MSM for bivariate parallel data, with states ‘both alive’, ‘individual 1 dead’, ‘individual 2 dead’ and ‘both dead’.

![Figure 3: The bivariate model for bivariate parallel data.](image)

The MSM provides a comprehensive view of the disease process, giving an efficient use of the incomplete information, when portions of the history of an
individual’s illness are known. In this framework, the so-called transition intensities provide the hazard for movement out of one state into another. These functions can also be used to determine the mean sojourn time in a given state of illness, the number of individuals in different states at a certain moment, and survival proportions in each state. Covariates in transition intensities can also explain differences in the course of the illness among the population. In this way, MSMs can reveal that different covariates affect different transitions, which would not be possible with other models, like, for example, Cox regression model with fixed covariates. 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. In fact, it is very unlikely that the risk of death in patients that received distinct treatments, for example, would be the same. Furthermore, the prognostic factors associated with the risk of death can be different in these groups of patients.

In medical applications, measurements of a disease are often made at several times, providing incomplete observations in some way. Because the process cannot be observed over an infinite time period, at times it will often be ended before an absorbing state is reached. In these situations the whole trajectory of the process was not observed, providing in this way 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 time of these events is unknown. Frequently, patients are observed at intermittent follow-up visits (e.g., in cohort studies), at which time individual data and covariate information are collected, but the information from the periods between visits is not available. In such cases the transition times of movement between states are not exactly observed and the states occupied in between visits are not known; these observations are said to be interval-censored. Also
frequent is the presence of left-truncation when dealing with multi-state models. For example, a common occurrence in the illness-death model is that for one patient to be selected, he must be in the “Healthy” state.

Note that censoring and truncation are two different concepts. Censoring indicates that the information about the process is only partially observed, whereas truncation represents an exclusion of a part of the population. In most cases, the truncation and censoring mechanisms are assumed independent from the process. In this paper we mainly discuss the analysis of data in which all the event times are observed exactly or right censored. In such cases, a crucial assumption is independent right-censoring. This (loosely) means that the individuals that are censored can be represented by those without censoring.

There exists an extensive literature on MSMs. Main contributions include books by Andersen et al. (1993) and Hougaard (2000). Recent reviews on this topic may be found in the papers Commenges (1999), Hougaard (1999) and Andersen and Keiding (2002). An issue of the Journal Statistical Methods in Medical Research, entirely devoted to these models, was published in 2002. Despite its potentialities, multi-state modelling is not used by practitioners as frequently as other survival analysis techniques. Lack of knowledge of the available software as well as misunderstanding of what multi-state modelling’s advantages rely on (compared to the simple Cox model), are probably responsible for this lack of popularity. We believe that the present paper contributes to fill an existing gap, by reviewing the most commonly used MSMs in an applied framework (Sections 2 and 4), and by reporting the ‘state of the art’ regarding the existing software (including a routine developed by the authors, Section 3). Benefits of using the multi-state approach can be readily understood from Section 4, in which two real data sets are considered and analyzed in great detail. One of these data sets (the
Galician breast cancer data is analyzed in this paper for the first time under the multi-state modelling perspective. In addition to the breast cancer dataset we also use the well-known and widely studied Stanford Heart Transplantation database. In both datasets, non-fatal events (recurrence and transplantation) are observed during the disease course. The analyses about the effect of these intermediate events on survival provide important information on the disease/recovery process.

Besides, this work includes a revision of one topic of major interest in the scope of MSMs, which has been recently investigated and which constitutes a new research area. This topic is that of the practical restrictions involved by using Markov MSMs. To this regard, we review some recent contributions and proposals for the estimation of transition probabilities which do not rely on the Markov assumption, see Section 2.

The article is organized as follows. Sections 2 focuses on the methodological aspects related to all the reviewed models, emphasizing on the estimation of the effects of covariates and estimation of transition probabilities. In Section 3, we present the available 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, a discussion of the article is given in Section 5.

2 Multi-state models

A multi-state process is a stochastic process $(X(t), t \in \mathcal{T})$ with a finite state space $S = \{1, \ldots, N\}$ and fulfilling some simplification assumptions. Here, $\mathcal{T} = [0, \tau]$, $\tau < \infty$ is the time interval of observation. This process has the information of the different transitions that occur to an individual over time, as well as the time at which these transitions take place. The process starts with the distribution of the initial state given by
\( p_j(0) = \mathbb{P}[X(0) = j], j \in S \). We shall call the probabilities, \( p_j(t) = \mathbb{P}[X(t) = j] \), state occupation probabilities. With the evolution of the process over time, a history \( \mathcal{X}_t \) (a \( \sigma \)-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. This multi-state process is fully characterized through transition probabilities between states \( h \) and \( j \), that we will express by:

\[
p_{bj}(s,t) = \mathbb{P}\left(X(t) = j\big| X(s) = h, \mathcal{X}_s \right)
\]

for \( h,j \in S, s,t \in \mathcal{S}, s \leq t \) or through transition intensities

\[
\alpha_{bj}(t) = \lim_{\Delta \to 0} \frac{p_{bj}(t,t+\Delta t)}{\Delta t}
\]

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

The process \( X(\cdot) \) is then governed by an intensity matrix \( Q(t) \) with \((h,j)\) entry \( \alpha_{bj}(t) \) or by a transition probability matrix, \( P(s,t) \) with \((h,j)\) entry \( p_{bj}(s,t) \). Using the notation of Commenges (2002), we then define the stochastic process as

\[
X(t) \sim \text{MSM}\left(\alpha_{bj}(t); h, j = 1, \ldots, N\right)
\tag{1}
\]

Various possible models for the transition rates between states can be accommodated in expression (1). The most common models are characterized through one of the following assumptions:

1. Time-Homogeneity: the intensities are constant over time, that is, the transition intensities are independent of \( t \) (Chiang, 1968; Kay, 1986).
2. **The Markov assumption**: future evolution only depends on the current state. That is, the transition intensities are independent of the history of the process (Cox and Miller, 1965; Chiang, 1968; Kay, 1986).

3. **The semi-Markov assumption**: future evolution not only depends on the current state, but also on \( t_h \), the entry time into that same state \( h \). Under this assumption we will consider intensity functions of the general form \( \alpha_{hj}(t,t-t_h) \) or, as the special homogeneous case \( \alpha_{hj}(t-t_h) \) (e.g., Andersen et al., 2000).

### 2.1 Markov models

Because of their simplicity, Markov models are the most used in the literature. In these models, transition probabilities and transition intensities depend on the individual’s past history only though the current state. The process \( (X(t), t \geq 0) \) is Markovian, if for any \( s, t \) with \( 0 \leq s < t \), and \( h, j, x(u) \in \{1,2,\ldots,N\} \), we have

\[
P[X(t) = j | X(s) = h, X(u) = x(u), 0 \leq u < s ] = P[X(t) = j | X(s) = h].
\]

Thus, the future of the process after time \( s \) depends only on the state occupied at time \( s \).

In Markov models, the transition probability matrix can be calculated from the intensity matrix, \( Q \), by solving the forward Kolmogorov differential equation (Cox and Miller, 1965),

\[
\frac{dP(s,t)}{dt} = P(s,t)Q(t).
\]

For example, for the illness-death model the transition probabilities are expressed as (all the rest can be deduced from these ones):

\[
p_{11}(s,t) = e^{-(A_{11}(s,t) + A_{21}(s,t))}, \quad (2)
\]

\[
p_{22}(s,t) = e^{-A_{22}(s,t)}, \quad (3)
\]
\[
p_{12}(s,t) = \int_s^t p_{11}(s,u)A_{12}(u)p_{22}(u,t)du,
\]

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

The survival function, \( S(t) \), is now expressed as,

\[
S(t) = p_{11}(0,t) + p_{12}(0,t).
\]

In some cases it can be very useful to study the conditional survival function, \( S(t \mid t_0) \), representing the probability of survival to time \( t \) given that the individual is “alive” at time \( t_0 \leq t \), expressed as \( S(t \mid t_0) = S(t)/S(t_0) \).

We shall now review several multi-state Markov models.

### 2.1.1 Time-homogeneous model

In homogeneous Markov models (HMM), all transition intensities are assumed to be constant as functions of time. Therefore, each transition probability, \( p_{ij}(s,t) \), depends only on \( t - s \), that is, \( p_{ij}(s,t) = p_{ij}(0,t - s) \). To simplify notation, we will use a matrix depending on only one argument in time \( p_{ij}(t - s) = p_{ij}(0,t - s) \). Then, subjected to \( P(0) = I \), the transition probabilities can be simply expressed in terms of the transition intensities, through the Kolmogorov relation \( P(t) = \exp(tQ) \). The matrix exponential can be calculated using the decomposition of \( Q \) into eigenvalues and eigenvectors. If \( Q \) has unique eigenvalues \( \nu_1, ..., \nu_k \), denoted as a vector of \( V \), and if \( U \) is the \((k-1) \times (k-1)\) matrix with the corresponding eigenvectors as its columns, then the solution is given by \( P(t) = \exp(tQ) = U \text{diag} \left( \exp(\nu_1 t), ..., \exp(\nu_k t) \right) U^{-1} \). See, for example, Cox and Miller (1965) for a proof.
Solutions for equations (2)-(4) are now the following:

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

Inference in these models is conducted through a general likelihood. We shall come to the construction of this likelihood in Section 2.1.3.

2.1.2 Piecewise constant intensities model

In some applications, the hypothesis of homogeneity may be unrealistic since, as time goes on, the illness tends to evolve. In these occasions, a non-homogeneous model is then recommended. Several non-parametric approaches to non-homogeneous processes have been proposed in the literature (Aalen and Johansen, 1978; Frydman, 1995; Joly et al., 2002). An alternative (parametric) procedure consists of partitioning the whole study period in two or more intervals, and then to fit a piecewise constant intensities model (Pérez-Ocón et al., 2001; Saint-Pierre et al., 2003), leading to transition intensity function as step functions. Here we consider a process where the transition intensities are defined by stepwise constant functions of type:

\[ \alpha_{lj}(t) = \alpha_{lj}, \quad \theta_{l-1} < t \leq \theta_{l}, \quad l = 1, 2, 3, ..., k \]

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

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 \( \chi^2_{k-r} \) distribution, where \( r \) is the number of parameters under \( H_0 \) and \( k \) 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 (1985).
There are limitations with the use of a piecewise model. For example, it is not always clear how best to choose the number of cutpoints and their appropriate values. The possibility of an arbitrary and/or opportunistic choice of cutpoints can be considered as a disadvantage of this method.

### 2.1.3 Likelihood construction

The inference for the time-homogeneous model is conducted by a general likelihood, which is derived as follows. Assume that the stochastic process \( X_t(\cdot) \) (with constant intensities) is observed at times \( t_{i,0} < t_{i,1} < \ldots < t_{i,m_i} \), where \( i = 1, \ldots, n \) are the indexed individuals, let us consider a general multi-state model, with a pair of states consecutively observed \((X_i(t_{i,r}), X_i(t_{i,r+1}))\). The general likelihood is then the product of all the terms over all the individuals and all the transitions (Kay, 1986),

\[
L = \prod_{i=1}^{n} \prod_{r=0}^{m_i-1} l_{i,r}
\]

(7)

where \( l_{i,r} = p_{X_i(t_{i,r}), X_i(t_{i,r+1})}(t_{i,r+1} - t_{i,r}) \) is the contribution to the likelihood for the \( i \)th individual for the pair of states \((X_i(t_{i,r}), X_i(t_{i,r+1}))\) observed.

The likelihood function for the piecewise constant model is built as follows. Let us assume \( k-1 \) cutpoints: \( 0 = \theta_0, \theta_1, \theta_2, \ldots, \theta_k = \infty \) and define one piecewise constant intensity matrix:

\[
Q(t) = Q_{l} \ , \ \theta_{l-1} \leq t < \theta_{l} \quad l = 1, 2, 3, \ldots, k
\]

Assuming that the process \( X_t(\cdot) \) is observed at times \( t_{i,0} < t_{i,1} < \ldots < t_{i,m_i} \), the likelihood is expressed as in (7). Let us define the following intervals: \( B_q = [\theta_q, \theta_{q+1}] \)
and \( C_s = [\theta_s, \theta_{s+1}] \) with \( q, s = 1, 2, \ldots, k - 1 \). Then, each contribution \( l_{i,r} \) is constructed as follows:

1. if \( t_{i,r} \in B_q \) and \( t_{i,r+1} \in C_q \), then
   \[
   l_{i,r} = p_{x_{i(r),x_{i(r+1)}}}^{Q_{q+1}}(t_{r+1} - t_r);
   \]

2. if \( t_{i,r} \in B_q \) and \( t_{i,r+1} \in C_{q+1} \), then
   \[
   l_{i,r} = p_{x_{i(r),x_{i(r+1)}}}^{Q_{q+1}}(t_{r+1} - t_r) \times p_{x_{i(r),x_{i(r+1)}}}^{Q_{q+2}}(t_{r+1} - q_{q+1});
   \]

3. if \( t_{i,r} \in B_q \) and \( t_{i,r+1} \in C_l \) with \( l - q \geq 2 \), then
   \[
   l_{i,r} = p_{x_{i(r),x_{i(r+1)}}}^{Q_{q+1}}(t_{r+1} - t_r) \times \left( \prod_{u=q+1}^{l-1} p_{x_{i(r),x_{i(r+1)}}}^{Q_{u+1}}(t_{r+1} - q_u) \right) \times p_{x_{i(r),x_{i(r+1)}}}^{Q_{l+1}}(t_{r+1} - q_{l+1}).
   \]

where \( P_{Q_l}^{l} \) is the transition probability matrix calculated from the intensity matrix \( Q_l \), \( l = 1, 2, 3, \ldots, k \).

If the process is continuously observed then the exact transition times are all known and in this case explicit estimators for all \( \hat{\alpha}_h^l \) are available. These are given by

\[
\hat{\alpha}_h^l = \frac{m_{hj}^l}{T_h^l},
\]

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 \( l \).

### 2.1.4 Nonparametric Model

Another non-homogeneous Markovian approach is the simple case of a nonparametric model. This approach can be thought as the generalization of the Kaplan-Meier estimate of the simple mortality model. For survival data, the transition probability from the ‘alive’ state into the ‘dead’ state may be estimated using the Kaplan-Meier estimator.
Aalen and Johansen (1978) propose a generalization of this approach to general MSMs with a finite number of states. Such a generalization is the so-called Aalen-Johansen estimator.

The Aalen-Johansen estimator can be defined for any Markov model, however, for simplicity, we assume the illness-death model. Assume that we have a sample of \( n \) continuously observed subjects with disease and death times, \( t_1, t_2, \ldots, t_n \). If we now assume that we have \( k \) events and \( n - k \) censored observations, then we may write \( t_{(1)} < t_{(2)} < \ldots < t_{(k)} \) for the \( k \) event times arranged in increased order. Let now \( n_{i} \) and \( n_{2i} \) denote the number of healthy and diseased subjects, respectively, just prior to the event time \( t_{(i)} \). Further, let \( d_{12i} \) be the number of subjects who become diseased at time \( t_{(i)} \), while \( d_{13i} \) and \( d_{23i} \) denote, respectively, the numbers of healthy and diseased subjects who die at that same time. Then transition probabilities (2)-(4) may be estimated by

\[
\hat{p}_{11}(s,t) = \prod_{s<t_{(i)}\leq t} \left(1 - \frac{d_{12i} + d_{13i}}{n_{i}}\right) \tag{8}
\]

\[
\hat{p}_{22}(s,t) = \prod_{s<t_{(i)}\leq t} \left(1 - \frac{d_{23i}}{n_{2i}}\right) \tag{9}
\]

\[
\hat{p}_{12}(s,t) = \sum_{s<t_{(i)}\leq t} \hat{p}_{11}(s,t_{(i-1)}) \frac{d_{12i}}{n_{i}} \hat{p}_{22}(t_{(i)},t) \tag{10}
\]

Because (8) and (9) are Kaplan-Meier estimators, we may use Greenwood’s formula to achieve a variance estimator for such transition probabilities, whereas, a variance estimator for (10) is given by
\[
\hat{\text{var}}(\hat{p}_{12}(s,t)) = \sum_{s < t_{i(1-1)}} p_{11}(s,t_{i(1-1)})^2 \left[ \hat{p}_{11}(t_{(i)},t) - \hat{p}_{12}(t_{(i)},t) \right]^2 \frac{n_{1i} - 1}{n_{1i}^3} d_{12i} \\
+ \sum_{s < t_{i(1-1)}} \left[ \hat{p}_{11}(s,t_{i(1-1)}) \hat{p}_{12}(t_{(i)},t) \right]^2 \frac{n_{1i} - 1}{n_{1i}^3} d_{13i} \\
+ \sum_{s < t_{i(1-1)}} \left[ \hat{p}_{12}(s,t_{i(1-1)}) \hat{p}_{22}(t_{(i)},t) \right]^2 \frac{n_{2i} - 1}{n_{2i}^3} d_{23i}
\]

For general cases, explicit expressions for the transition probabilities like those mentioned above cannot be given. Further details can be found in Andersen et al. (1993) and Keiding et al., (2001).

2.2 Non-Markov models

Traditionally, statistical methods for analyzing MSMs depend on the Markov assumption. Under the Markov assumption, the transition intensities depend on the current time and the state currently occupied; they do not depend on the patient history (length of stay in the current state; time of transition to the current state; states previously visited; patient characteristics measured before, etc.). By ignoring the disease history behaviour, these models may carry severe limitations which can make the model inappropriate. It is a fact that the future health of recently diseased individuals may be different from those who have been diseased for a long time.

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. (2001) and Datta and Satten (2001) studied the performance of the Aalen-Johansen estimator of stage occupancy probabilities when the process is not Markovian. These authors established the consistency of Aalen-Johansen estimators of the occupation probabilities in a non-Markov process under independent censoring. Later, Glidden (2002) developed robust confidence bands for those event curves. Recently, Meira-Machado et al., (2006), verified that in non-Markov situations, the use of Aalen-Johansen estimators to empirically estimate the transition probabilities, \( p_{ij}(s,t) \), may be inappropriate. These authors propose, in the scope of the illness-death model, alternative “Markov-free” estimators for the transition probabilities which do not rely on the Markov assumption. The proposed estimators are expressed as,

\[
\hat{p}_{11}(s,t) = \frac{1 - \hat{H}(t)}{1 - \hat{H}(s)},
\]

(11)

\[
\hat{p}_{12}(s,t) = \frac{1}{1 - \hat{H}(s)} \sum_{i=1}^{n} W_i \phi_{s,t} (Z_{[i]}, Y_{(i)}),
\]

(12)

\[
\hat{p}_{22}(s,t) = \sum_{i=1}^{n} W_i \hat{\phi}_{s,s} (Z_{[i]}, Y_{(i)}) \bigg/ \sum_{i=1}^{n} W_i \hat{\phi}_{s,s} (Z_{[i]}, Y_{(i)}),
\]

(13)

where \( Y = \min(T, C) \) (\( T \) is the survival time of the process and \( C \) the right-censoring variable which is assumed to be independent of the process), \( W_i \) are the Kaplan-Meier weights attached to \( Y_{(i)} \), and \( Z \) is the sojourn time in state 1, with distribution function denoted by \( H \), and \( \hat{H} \) its Kaplan-Meier estimator; \( \phi_{s,t}(u,v) = \mathbb{I}(s < u \leq t, v > t) \) and
\( \tilde{\phi}_{s,t}(u,v) = \mathbb{I}(u \leq s, v > t) \). In these expressions, \( Y^{(1)} \leq \cdots \leq Y^{(o)} \) denote the ordered sample of the \( Y_i \)'s, and \( Z^{(i)} \) for the pair attached (concomitant) to the \( Y^{(i)} \) value.

Basic ideas behind estimators (11)-(13) 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. Since the methods in Meira-Machado et al. (2006) are free of the Markov assumption, it could be used to introduce such tests (at least in the scope of the illness-death model) by measuring their discrepancy with respect to Markovian estimators. This topic is currently under investigation.

Recall that for the illness-death model, the survival function is expressed as the sum of two occupation probabilities (see (5)). Besides, the Aalen-Johansen estimators of \( p_{11}(0,t) \) and \( p_{12}(0,t) \) are consistent regardless the Markov assumption (see Datta and Satten, 2001). Then, the survival estimate is consistent whether we use Aalen-Johansen estimators or “Markov-free” estimators. The same argument is valid for the conditional survival function \( S(t \mid t_0) \). The survival prognosis of an individual being in state 2 at time \( t_0 \), denoted by \( S(t \mid X_{t_0} = X(u), u \in [0, t_0]; X(t_0) = 2) \), is given by \( p_{22}(t_0, t) \). As mentioned above, in non-Markov situations, the use of Aalen-Johansen estimators to estimate this transition probability may be inappropriate. In such cases, flexible estimation methods are recommended.

### 2.3 Multi-state regression models

Defining \( X_j(t) \) 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 with the intensity rates through a covariate vector, $Z$, possibly time-dependent. For a general regression model we can write

$$\alpha_{hji} (\cdot) = \varphi(\alpha_{hj0} (\cdot), \beta_{hj}^T Z_i)$$

where $\alpha_{hj0} (\cdot)$ is the baseline intensity function between states $h$ and $j$, $\beta_{hj}$ is the vector of regression parameters, and $Z_i$ is the covariate vector for subject $i$.

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

$$\alpha_{hji} (t; Z) = \alpha_{hj0} (t) \exp \left( \beta_{hj}^T Z \right)$$

(14)

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,

$$\alpha_{hji} (t; Z) = \alpha_{hj0} (t) + \beta_{hj}^T \left( t \right) Z .$$

(15)

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

$$\alpha_{hj}(Z) = \alpha_{hj} \exp \left( \beta_{hj}^T Z \right)$$

(16)

is traditionally assumed to relate the transition intensities $\alpha_{hj}$ with covariates $Z$. The introduction of covariates in these models may bring out important biological insights in the disease process. Furthermore, it allows prediction of probabilities tailored to individual patients. In both cases (HMM and NHM) the inference may be based on the likelihood (7) by replacing the transition intensities $\alpha_{hj}$ by those given above in equation (16).
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

$$\alpha_{hj}(t; Z) = \alpha_{hj0}(t)\exp(\beta_{hj}^TZ)$$

assuming the process to be Markovian. These models are known as Cox Markov models (CMM).

For the hazard of ‘death’ without disease, $\alpha_{13}(t; Z)$, survival times from patients that experienced the disease are taken as censored in 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 the beginning of the disease. Survival times of patients who did not become diseased are taken 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.

Note that in some cases we may assume some conditions about 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

$$\alpha_{13}(t; Z) = \alpha_{130}(t)\exp(\beta_{13}^TZ) \text{ and } \alpha_{23}(t; Z) = \alpha_{230}(t)\exp(\beta_{23}^TZ + \delta).$$
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 (Andersen et al., 2000):

\[
\hat{p}_{11}(s,t|Z) = \prod_{s < u < t} \left( 1 - \sum_{j=2}^{\infty} d \bar{A}_{j}(u|Z) \right)
\]

\[
\hat{p}_{22}(s,t|Z) = \prod_{s < u < t} \left( 1 - d \bar{A}_{2}(u|Z) \right)
\]

\[
\hat{p}_{12}(s,t|Z) = \sum_{u < t} \hat{p}_{11}(s,u^{-}|Z) d \bar{A}_{2}(u|Z) \hat{p}_{22}(u^{+},t|Z).
\]

where \( \bar{A}_{hj}(t|Z) = \bar{A}_{hj0}(t) \exp(\beta_{hj}^T Z) \) is the estimate of the cumulative intensity function with \( \bar{A}_{hj0}(\cdot) \) the Breslow estimator for \( A_{hj0}(t) = \int_{0}^{t} \alpha_{hj0}(u) du \).

The conditional survival probability, \( S(t|Z) \), can now be estimated by

\[
\hat{S}(t|Z) = \hat{p}_{11}(0,t|Z) + \hat{p}_{12}(0,t|Z)
\]

with changes in its values at observed failure times.

One alternative approach is to use a semi-Markov assumption in which future of the process does not depend on the current time but rather on the duration in the current state. Homogeneous semi-Markov models (Andersen et al., 2000) 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. (2000) and Andersen and Keiding (2002).


2.3.2 Introducing flexible covariate effects

In multi-state survival analysis regression models are often used to relate the individual characteristics with intensity rates through a covariate vector, $Z$, possibly time-dependent. In this case, the Cox proportional hazards model (14) is typically assumed, the additive hazards model (15) being a possible alternative. In both approaches, the effect of prognostic factors is assumed to have a linear (or log-linear) functional form. However, if the effect is highly nonlinear, this assumption is violated leading to erroneous statistical conclusions: bias and decreased power of tests for statistical significance (Anderson and Fleming, 1995). In the Cox framework, the incorrect functional form for a covariate can also lead to a diagnosis of nonproportional hazards (Therneau and Grambsch, 2000). 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 paper. Software for fitting such methods is available as part of S-plus and R statistical languages. The issue of how to implement nonlinear 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 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 Hastie (1987) considered nonparametric 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 $f_j(\cdot), j = 1, \ldots, p$ are unspecified smooth covariate functions. These models have been studied by several authors using various nonparametric techniques. Spline based smoothing methods were considered, for instance, by O’Sullivan (1988), Hastie and Tibshirani (1990), Gray (1994) and Kooperberg, Stone and Truong (1995). Smoothing is achieved by imposing a penalty for curve roughness. These models have gained recent popularity due to Eilers and Marx (1996) when they introduced penalties to the B-splines. Huang et al. (2000) used functional ANOVA decompositions to study a general class of models that includes the additive proportional hazards model. Recently, Huang and Liu (2006) considered a proportional hazards model of the form

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

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

As mentioned above, these methods can be considered for Cox Markov (semi-Markov) models. Furthermore, they can be used to test the presence of nonlinear effects as well as to identify its correct covariate functional form. The introduction of these nonlinear effects on the transition rates can help to prevent specification errors as well as to provide important information about the relationship between prognostic factors and disease risk.
3 Existing software

While the time-dependent Cox model (TDCM) can be fitted through all the major statistical packages, MSMs need specialized software, most of which are written in FORTRAN, R or SAS. Marshall and Jones (1995) have developed a FORTRAN program called MARKOV, for fitting multi-state Markov models with constant transition intensities and covariates. Later, Alioum and Commenges (2001) presented a new computer program, called MKVPCI, which extends MARKOV by allowing piecewise-constant intensities with different values in at most three time intervals. Jackson and Sharples (2002) developed the R package *msm*, implementing several functions for fitting continuous-time Markov MSMs to categorical processes observed at arbitrary times. Recently, Paes and Lima (2004) developed a SAS macro, called PTRANSIT, for estimating transition probabilities in semi-parametric models for recurrent events. Hui-Min, Ming-Fang and Hsiu-His (2004) have developed a SAS macro for estimating the transition parameters in non-homogeneous (Weibull distributions, log-logistic, etc.) $k$-state progressive Markov model. Rosthøj, Andersen and Abildstrom (2004) developed two SAS macros for estimation of the cumulative incidence functions for competing risks survival data. Recently, Wangler et al., (2006) developed a R library called *changeLOS* that implements the Aalen-Johansen estimator for general multi-state models with nonparametric hazards.

Most of the existing software for MSMs presents, however, some difficulties and limitations in practice. In some clinical studies a model with the Markov assumption may be appropriate, while for others the semi-Markov is preferable. In several cases a homogeneous model will be satisfactory, while in others not. Furthermore, possible comparisons between different MSMs are rather difficult because each of the current programs requests its own input data-structure. In addition, most of the programs
available only provide the regression parameters estimates and do not supply graphical output for the survival estimates and for transition probabilities estimates.

We developed a user-friendly R library, that we called \textit{tdc msm} for the analysis of multi-state survival data, which is freely available upon request. Specifically, the new software may be used to fit the TDCM but also the others reviewed MSMs including various transient states and one 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 rate estimates for continuous covariates (Eilers and Marx, 1996). Moreover, users are able to include any number of covariates on transition intensities, whether the transition times are known or not. In this way, users may easily analyze the results offered by the various models in order to compare them and make decisions accordingly. The \textit{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. (2006).

Examples of analysis using this software will be shown using two datasets: the Stanford Heart Transplant Study and our own dataset on breast cancer.

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 article covers the period until April 1, 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 from the entry of the patient in the study (in days), and a vector of covariates including age at acceptance, year of acceptance, previous 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.

**Time-dependent Cox regression model**

One issue in the framework of the Cox model is the choice of time scale (e.g., Korn et al., 1997; Thiébaut and Bénichou, 2004; Pencina et al., 2006). 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 time-dependent Cox model (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 (1982). 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 \textit{start} and \textit{stop} mark the time interval \([\textit{start}, \textit{stop}]\) for the data, while \textit{event} is an indicator variable taking on value 1 if there was a death at time \textit{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 enrolment. Both subjects remain alive until the end of the follow-up. The time from enrolment to death for the third and fourth patients is 263 and 72 days, respectively, and the last patient receive a new heart at day 71. The data for these four patients displayed is represented as follows

<table>
<thead>
<tr>
<th>id</th>
<th>start</th>
<th>stop</th>
<th>event</th>
<th>transplant</th>
<th>age</th>
<th>year</th>
<th>surgery</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>33.2238</td>
<td>1.57426</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>1800</td>
<td>0</td>
<td>1</td>
<td>33.2238</td>
<td>1.57426</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>0</td>
<td>1401</td>
<td>0</td>
<td>0</td>
<td>30.5353</td>
<td>1.58248</td>
<td>0</td>
</tr>
<tr>
<td>27</td>
<td>0</td>
<td>263</td>
<td>1</td>
<td>0</td>
<td>8.7858</td>
<td>1.59069</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>0</td>
<td>71</td>
<td>0</td>
<td>0</td>
<td>54.0233</td>
<td>1.68378</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>71</td>
<td>72</td>
<td>1</td>
<td>1</td>
<td>54.0233</td>
<td>1.68378</td>
<td>0</td>
</tr>
</tbody>
</table>

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 \texttt{survival} library as follow:

\begin{verbatim}
> coxph(Surv(start, stop, event) ~ transplant + age + year + surgery, data = heart)
\end{verbatim}
Or using our own library *tdc.msm* (with the same input data file) through the following input command:

```r
> 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 Hu (1977) and Kalbfleisch and Prentice (1980). 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 analyzing 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-value > 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 yielded 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 (1977).
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 dataset, “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 3 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 often gives 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.

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, 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 summarized by the state occupied at time \( t \). The Markov assumption may be checked, among others, by including covariates in the modelling process (Kay, 1986). 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 (past) is important on the transition from the disease ("new heart") state to death. For doing that, let \( Z = "\text{time spent in state 1}" \) (alternatively, we could use "time since entry in state 2"), and \( t \) the current time. Fitting a model \( \alpha_{23}(t; Z) = \alpha_{230}(t) \exp(\beta Z) \), we now need to test \( \beta = 0 \), i.e., 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, using time since entry in study as the time scale, the effect of time spent in state 1 is not significant (p-value > 0.05) indicating that the Markov model is satisfactory for the Stanford data.

In the following we shall assume the time-scale to be the time-on-study.

**Cox Markov Model**

These models can be fitted through most of the statistical packages as long as we use a counting process notation, representing each patient by several observations (Therneau and Grambsch, 2000). 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:

<table>
<thead>
<tr>
<th>id</th>
<th>start</th>
<th>stop</th>
<th>event</th>
<th>transplant</th>
<th>age</th>
<th>year</th>
<th>surgery</th>
<th>transition</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>33.2238</td>
<td>1.57426</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>0</td>
<td>25</td>
<td>1</td>
<td>0</td>
<td>33.2238</td>
<td>1.57426</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>1800</td>
<td>0</td>
<td>1</td>
<td>33.2238</td>
<td>1.57426</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>26</td>
<td>0</td>
<td>1401</td>
<td>0</td>
<td>0</td>
<td>30.5353</td>
<td>1.58248</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>26</td>
<td>0</td>
<td>1401</td>
<td>0</td>
<td>0</td>
<td>30.5353</td>
<td>1.58248</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>27</td>
<td>0</td>
<td>263</td>
<td>1</td>
<td>0</td>
<td>8.7858</td>
<td>1.59069</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>0</td>
<td>263</td>
<td>0</td>
<td>0</td>
<td>8.7858</td>
<td>1.59069</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>28</td>
<td>0</td>
<td>71</td>
<td>0</td>
<td>0</td>
<td>54.0233</td>
<td>1.68378</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>0</td>
<td>71</td>
<td>1</td>
<td>0</td>
<td>54.0233</td>
<td>1.68378</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>28</td>
<td>71</td>
<td>72</td>
<td>1</td>
<td>1</td>
<td>54.0233</td>
<td>1.68378</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

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 “\textit{survival}” library and our library “\textit{tdc.msm}”. While the \textit{survival} library requires some extra effort with respect to data preparation (the long format data-structure shown above), the \textit{tdc.msm} library use the same dataset previously used for Cox time-dependent analyses. We can use Terry Therneau’s library as follow (using the new dataset 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 \texttt{tdc.msm} the following input command provides the (same) results for all the 
transitions

> \texttt{tdc.msm(heart, formula=c(6,7,8), models= “CMM”)}

Furthermore, this output performs also a test for checking the Markov assumption (as 
discussed before). The results obtained from fitting this model showed 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 $\alpha_{13}(t)$ (HR:0.753; 95\%CI: 0.606-0.936). Age 
revealed to be the best predictor for the mortality transition $\alpha_{23}(t)$ of transplanted 
patients (HR:1.050;95\%CI:1.008–1.100), and also for transition $\alpha_{12}(t)$, corresponding 
to receive a new heart (HR:1.032;95\%CI:1.004-1.060). The effect of age on the 
mortality intensity in patients without transplant was not significant 
(HR:1.020;95\%CI:0.984–1.057). No significant effect of a previous surgery was 
found.

\textit{Homogeneous Markov Model}

Like the preceding studied models, the HMM 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 (see Table 1) are in good agreement with those obtained through the CMM. Some exceptions are the surgery effect, which was not a significant factor in any of the previously studied models, and now reveals a significant effect on survival for transplanted patients (HR: 0.306; 95% CI: 0.128 – 0.730). 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: 0.739; 95% CI: 0.595 – 0.919).

These results were obtained using the following input command:

\[ \text{tdc.msm(heart, formula=c(6,7,8), models= "HMM")} \]

Also in the HMM 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 1 that the fitted HMM 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 1.163, 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 then 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 were used to test if the regression parameters are statistically different from zero. This statistic has an approximately \( \chi^2 \) distribution under \( H_0 : \beta_{ij} = 0 \). Estimated parameters for the simplified model (results not shown) did not differ substantially from those obtained by the initial model (shown in Table 1).
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 \( \chi^2 \)) yields a value of 18.5, suggesting that the transplant is significantly associated to a diminishing in mortality risk. Note again that likelihood ratio tests can also be used for constructing a test of \( H_0 \).

Table 1. Multi-state homogeneous Markov model. Estimated transition rates and hazard rates. Stanford Heart Transplantation Data.

| TR (SE) | 0.0137 (0.0017) | 0.0054 (0.0011) | 0.0018 (0.0003) |
| HR (95%CI) | Age |
| 1 \( \rightarrow \) 2 | 1.068 (1.039 – 1.098) |
| 1 \( \rightarrow \) 3 | 1.056 (1.020 – 1.093) |
| 2 \( \rightarrow \) 3 | 1.076 (1.030 – 1.125) |
| Year |
| 1 \( \rightarrow \) 2 | 0.975 (0.852 – 1.116) |
| 1 \( \rightarrow \) 3 | 0.739 (0.595 – 0.919) |
| 2 \( \rightarrow \) 3 | 1.109 (0.928 – 1.325) |
| Surgery |
| 1 \( \rightarrow \) 2 | 1.368 (0.737 – 2.539) |
| 1 \( \rightarrow \) 3 | 0.959 (0.277 – 3.315) |
| 2 \( \rightarrow \) 3 | 0.306 (0.128 – 0.730) |

TR=Transition rate; SE=Standard error; HR=Hazard ratio; CI=Confidence Interval.

The goodness-of-fit of a MSM can be assessed by comparing the observed and predicted number of patients undergoing each transition. For HMMs this can also be done studying baseline hazards for CMMs. Results (not reported here) showed 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. 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, Kay (1986) 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 (see section 2.2.1.2). Such a model will now be constructed.

*Piecewise constant intensities model*

In this section we build a piecewise constant intensities model like (6) 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 $\theta$, a value of $\theta = 80$ days was selected, and two intervals ($\text{time} \leq 80\text{ days, time} > 80\text{ days}$) were then considered. Again, we mention that 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 falls in all intervals.

Focusing mainly on the short-term survival period of 80 days, we verify that for the non-homogeneous model, the agreement between predicted and observed percentages of patients in each transition is globally satisfactory, being clearly better than that obtained previously using a homogeneous model.

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 was fitted excluding this covariate

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

From this new model we verified that in both intervals the resulting estimates for the mortality intensity were lower in transplanted patients. When examining the fixed covariate effects, we see that, for time $\leq 80$ days, age at acceptance is a significant predictor in transitions from state 2 to state 3 (HR: $1.057; 95\%$CI: $1.000-1.116$), while the effect of year is only significant on the mortality intensity in patients without transplant (HR: $0.737; 95\%$CI: $0.587-0.927$). For the second interval (time $>80$ days), however, the only significant covariate was age at acceptance for the transition from state 1 to state 2 (HR: $1.233; 95\%$CI: $1.050-1.449$).

Several reviewed methods of Section 2.1 were illustrated using the Stanford data, providing some guidance about the use of these methodologies for studying the course of the illness. This dataset has been widely studied in the medical literature focusing on the effect of the covariates. Among others, Turnbull, Brown and Hu (1974), Mantel and Byar (1974) and Crowley and Hu (1977) studied the Stanford Heart data reaching a small negative influence of transplantation on hazard, but without statistical significance. In contrast, using multi-state models we verified that the transplant is significantly associated to a diminishing in mortality risk. The same conclusion was obtained using a Cox model with age as the time-scale. Multi-state modeling also showed 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 in these studies is to identify which prognostic factors are predisposing to recurrence. Furthermore, it is important to investigate how these factors behave in the disease process. The occurrence of recurrence affects the patient 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. 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 its 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, tumour 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) revealed to be important prognostic factors of survival. None of the treatments obtained a statistical significant effect on survival.

Again, the Markov assumption can be checked by including covariates in the modelling process. For this application we verified 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 to weaken the often used Markov assumption. 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 follow

\[
> \text{coxph(Surv(stop-start, event) ~ perimet + size + age + radio + chemo + hormone, } \\
data = \text{breast, subset = (transition == 1))}
\]

\[
> \text{coxph(Surv(stop-start, event) ~ perimet + size + age + radio + chemo + hormone, } \\
data = \text{breast, subset = (transition == 2))}
\]

The same results are obtained using library \texttt{tdc.msm} with

\[
> \text{tdc.msm(breast, formula=c(6,7,8,9,10,11), models= "CSMM")}
\]

Table 2 summarizes results from fitting such a model, showing that some of the significant covariates in the simple Cox model turn out to affect the transition intensities differently. 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 \( \alpha_{13}(\cdot) \) is opposite to that on mortality transition. Results also indicate a significant effect of chemotherapy on recurrence transition.
Table 2: Estimated effects in Cox models for the recurrence intensity, \(\alpha_{12}(t)\), and for the mortality, \(\alpha_{23}(t - T_{12})\), after recurrence in the breast cancer data.

<table>
<thead>
<tr>
<th>Transitions</th>
<th>Covariate</th>
<th>(\hat{\beta})</th>
<th>SE</th>
<th>HR</th>
<th>95% CI for HR</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tumour size</td>
<td>0.099</td>
<td>0.023</td>
<td>1.104</td>
<td>1.056 – 1.155</td>
<td>0.000</td>
</tr>
<tr>
<td>1 (\rightarrow) 2</td>
<td>Age</td>
<td>-0.0004</td>
<td>0.007</td>
<td>0.999</td>
<td>0.985 – 1.103</td>
<td>0.950</td>
</tr>
<tr>
<td></td>
<td>Perimeter</td>
<td>0.055</td>
<td>0.005</td>
<td>1.057</td>
<td>1.047 – 1.067</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Radiotherapy</td>
<td>0.001</td>
<td>0.208</td>
<td>1.001</td>
<td>0.665 – 1.506</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>Chemotherapy</td>
<td>-0.740</td>
<td>0.246</td>
<td>0.477</td>
<td>0.294 – 0.773</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>Hormonotherapy</td>
<td>0.206</td>
<td>0.213</td>
<td>1.229</td>
<td>0.809 – 1.867</td>
<td>0.330</td>
</tr>
<tr>
<td></td>
<td>Tumour size</td>
<td>0.078</td>
<td>0.026</td>
<td>1.081</td>
<td>1.026 – 1.140</td>
<td>0.003</td>
</tr>
<tr>
<td>2 (\rightarrow) 3</td>
<td>Age</td>
<td>0.024</td>
<td>0.009</td>
<td>1.024</td>
<td>1.006 – 1.040</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>Perimeter</td>
<td>0.026</td>
<td>0.007</td>
<td>1.026</td>
<td>1.012 – 1.040</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Radiotherapy</td>
<td>-0.355</td>
<td>0.249</td>
<td>0.701</td>
<td>0.430 – 1.140</td>
<td>0.150</td>
</tr>
<tr>
<td></td>
<td>Chemotherapy</td>
<td>0.162</td>
<td>0.278</td>
<td>1.176</td>
<td>0.681 – 2.030</td>
<td>0.560</td>
</tr>
<tr>
<td></td>
<td>Hormonotherapy</td>
<td>-0.027</td>
<td>0.272</td>
<td>0.973</td>
<td>0.571 – 1.660</td>
<td>0.920</td>
</tr>
</tbody>
</table>

SE = Standard Error, HR = Hazard Ratio; CI=Confidence Interval.

For continuous covariates it is possible to test whether the covariate effect is linear or not. Fitting a Cox model with penalized splines (Eilers and Marx, 1996), and using the number of degrees of freedom which minimizes AIC, revealed a nonlinear covariate effect for perimeter on the recurrence transition (p-value < 0.05). There is no evidence of nonlinear 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,

\[
> \text{tdc.msm(breast, formula=c(6,7,8,9,10,11), models= “CSMM”, graphcov = 1)}
\]

Figure 5 displays the smoothed log hazard for the covariate perimeter, obtained by using penalized splines together with the 95 percent pointwise confidence limits. This plot suggests that the log-hazard is rather constant until around 60 and then rapidly increase until 95, remaining nearly constant afterwards. This plot illustrates advantages from using graphical tools with flexible (nonlinear) relative risk form to understand the effects of prognostic factors on the recurrence incidence.
Figure 5. Log-hazard estimation with penalized splines for perimeter (mm) (with 95% pointwise confidence bands) for the recurrence transition. Breast cancer data.

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 described in Section 2.2. In Figure 6 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.
Figure 6. Estimated transition probabilities for Aalen-Johansen estimator (solid line) and non-Markov model (dashed line). Breast cancer data.
As it can be seen from Figure 6, 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. With regard to the survival prognosis, we observe serious departures between both survival curves for individuals who have had recurrence. Differences are clearer with the progression of time, showing that the prognosis using the “Markov-free” estimator is poorer than Aalen-Johansen prognosis.

5. Discussion

In this paper, we have illustrated the usefulness of multi-state models in the analysis of survival data. We have discussed the application of these methods using two datasets: the Stanford heart transplant study and a new dataset on breast cancer. We hope that this discussion will encourage the applied researchers to use multi-state modelling more frequently or with greater confidence, as part of their routine data analysis techniques. Importantly, this article includes an up-to-date review of the existing software as well as of some modern, flexible methods to cope with nonlinear covariate effects and with non-Markov data structures. The potential advantages of these methods were illustrated through the referred datasets.

When analyzing 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 time-dependent Cox model we have verified that the multi-state modelling yielded new insights while confirming some of the results obtained with the time-dependent Cox model.
We also used a new dataset 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 verify that, on the recurrence transition, some covariate effects are best represented by smooth, nonlinear 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 a model for the marginal survival distribution. 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 but rather as supplements that offer additional information.

Acknowledgments

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 and PGIDIT06PXIC300117PN. Per Kragh Andersen was supported by a grant (R01 CA54706-12) from the National Cancer Institute. We are also grateful to Professor Puente Dominguez (in memoriam) and to Dr. Elisabeth Chávez for the use of the breast cancer data.

References


Huang JZ, Liu L. Polynomial Spline Estimation and Inference of Proportional Hazards Regression Models with Flexible Relative Risk Form. *Biometrics* 2006;


Department of Biostatistics
University of Copenhagen
Øster Farimagsgade 5
P.O. Box 2099
1014 Copenhagen K
Denmark

06/1 Carstensen, B. Demography and epidemiology: Age-Period-Cohort models in the computer age.

06/2 Carstensen, B. Demography and epidemiology: Practical use of the Lexis diagram in the computer age or: who needs the Cox-model anyway?


06/4 Christensen, K.B. & Kreiner, S. A Monte Carlo approach to unidimensionality testing in polytomous Rasch models.

06/5 Keiding, N. Event history analysis and the cross-section.

06/6 Ditlevsen, S.D. & Lansky, P. Estimation of the input parameters in the Feller neuronal model.

06/7 Kvist, K., Andersen, P.K. & Kessing, L.V. Repeated events and total time on test.


06/9 Budtz-Jørgensen, E. Estimation of the benchmark dose by structural equation models.

06/10 Christensen, K.B., Feveille, H., Kreiner, S. & Bjørner, J.B. Adjusting for mode of administration effect in surveys using mailed questionnaire and telephone interview data.

06/11 Dalgaard, P. New R functions for multivariate analysis.


06/13 Kvist, K., Harhoff, M.G., Andersen, P.K. & Kessing, L.V. Non-parametric estimation and model checking procedures for marginal gap time distributions for recurrent events.

07/1 Meira-Machado, L., de Uña-Álvarez, Cardarso-Suárez, C. & Andersen, P.K. Multi-state models for the analysis of time to event data.