Inference in the progressive three-state model

Luís Meira-Machado, Carmen Cadarso-Suárez, Jacobo de Uña-Álvarez

***Abstract*—** In longitudinal studies of disease, patients can experience several events through a follow-up period. In these studies, the sequentially ordered events (gap times) are often of interest. The events of concern may be of the same nature (e.g. cancer patients may experience recurrent disease episodes) or represent different states in the disease process (e.g. alive and disease-free, alive with recurrence and dead). If the events are of the same nature this are usually referred as recurrent event, whereas if they represent different states (i.e. multi-state models) they are usually modeled thought their intensity functions. In this paper we present nonparametric estimators for several quantities in a progressive three-state model. We present a simple estimator for the bivariate distribution function for censored gap times and estimators for the transition probabilities. The proposed methods can be easily extended for the progressive *k*-state model (with a vector of *k* gap times). Another major goal is to study the relationship between the different covariates and disease evolution. The proposed methods were applied to a database on breast cancer from Galicia, Spain. Software (in R) implementing the methods proposed in this paper were developed by the authors.

***Keywords*—**bivariate censoring, Kaplan-Meier, Multi-state model, Proportional Hazards Model.

I. Introduction

In many medical studies, patients may experience several events. The analysis in such studies is often performed using multi-state models [1]-[4]. These models are very useful for describing event history data offering a better understanding of the process of the illness, and 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.

The complexity of a multi-state model greatly depends on the number of states defined and by the transitions allowed between these states. The simplest form of multi-state model is the "two-state" model, or *mortality model*, 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*, illustrated in Figure 1. It has three states and the only possible transitions are 1→2 and 2→3. Note that for the progressive three-state model we may assume that the transition intensity from state 2 into state 3 might depend, in some way, on the entry time in state 2, denoted by t_{12}.

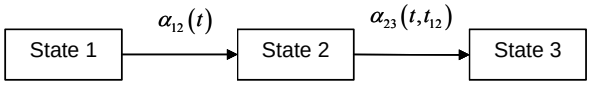

Fig. 1 Progressive three-state model

The scope of multi-state models provides a rich framework to handle complex situations involving more than two states and a number of possible transitions among then. The most common models in literature include: the progressive *k*-state model (which is a generalization of the mortality model and the progressive three-state model), depicted in Figure 2; the illness-death model, depicted in Figure 3; the bivariate model and the competing risks model (not shown).

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, are probably responsible for this lack of popularity. We believe that the present paper contributes to fill an existing gap by presenting new estimation methods for several functions of interest in medical studies. The proposed methods can be implemented using software developed by the authors. Specifically, we focus on the three-state model of Figure 1. In this model, the times between consecutive events (which define states 2 and 3) are often of interest. In the scope of this model we present a nonparametric estimator of the bivariate distribution function of the gap times. Some related problems as the estimation of the marginal distribution of the second gap time will be discussed. Estimators for the transition probabilities are also given.

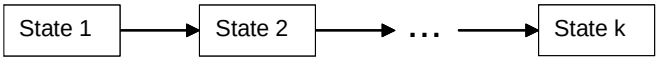

Fig. 2 Progressive k-state model

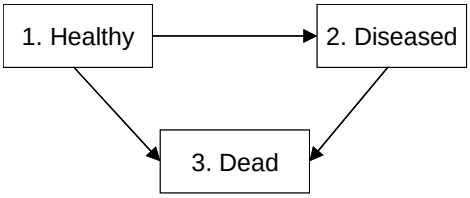

Fig. 3 Illness-death model

The layout of this paper is organized as follows: notation is introduced and nonparametric estimators for bivariate distribution function and transition probabilities are presented in Section 2; regression methods are discussed in Section 3, focusing on the proportional hazards model; an example of

application to a real breast cancer data set is presented in Section 4. Finally, we conclude with a discussion section.

II. NONPARAMETRIC ESTIMATORS

In this section we present nonparametric estimators for the progressive three-state model. We review some of the methods proposed in the literature and propose new estimators for several functions. The idea behind the new estimators proposed below is using the Kaplan-Meier estimator pertaining to the distribution of the total time to weight the data.

Multi-state process are characterized through: transition probabilities between states *h* and *j* that we express as

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

$s \le t$, where H_{s-} $(\sigma - algebra)$denotes the history of the process and is generated consisting of the observation of the process over the interval $[0, s[$; or through transition intensities that we express as

$$\alpha_{hj}(t) = \lim_{\Delta t \to 0} p_{hj}(t, t+\Delta t)/\Delta t$$

representing the instantaneous hazard of progression to state *j* conditionally on occupying state *h*, and that we shall assume exist.

A number of possible models for the transition rates have been studied. These include:

(a) *Time-homogeneity*: the intensities are constant over time, that is, independent of *t*;

(b) *The Markov assumption*: the transition intensities only depend on the history of the process through the current state;

(c) *The semi-Markov assumption*: 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 $\alpha_{hj}(t-t_h)$.

Consider the progressive three-state model depicted in Figure 1. Let $\{X(t), t \ge 0, X(0) = 1\}$ denote the underlying stochastic process where $X(t)$ denote the state being occupied at time *t*, for which all individuals are in state 1 at time zero. We represent the stochastic behaviour of the process by a random vector (T_{12}, T_{23}), where T_{hj} is the potential transition from state *h* to state *j*, $1 \le h < j \le 3$. The pair (T_{12}, T_{23}) is a pair of gap times of successive events, which are observed subjected to random right-censoring. Let *C* be the right-censoring variable, assumed to be independent of (T_{12}, T_{23}) and let $Y = T_{12} + T_{23}$ be the total time. Because of this, we only observe $\left(\tilde{T}_{12i}, \tilde{T}_{23i}, \Delta_{1i}, \Delta_{2i}\right)$, $1 \le i \le n$, which are *n* independent replications of $\left(\tilde{T}_{12}, \tilde{T}_{23}, \Delta_1, \Delta_2\right)$, where $\tilde{T}_{12} = T_{12} \wedge C$, $\Delta_1 = I(T_{12} \le C)$, and $\tilde{T}_{23} = T_{23} \wedge C_2$, $\Delta_2 = I(T_{23} \le C_2)$, with $C_2 = (C - T_{12})I(T_{12} \le C)$ the censoring variable of the second gap time. Note that $\Delta_2 = 1$ implies $\Delta_1 = 1$. Hence $\Delta_2 = I(Y \le C)$. Define $\tilde{Y} = Y \wedge C$ and let *H* and *G* denote the distribution functions of T_{12} and *C*, respectively. Since T_{12} and C are independent, the Kaplan-Meier product-limit estimator based on the pairs $\left(\tilde{T}_{12i}, \Delta_{1i}\right)$'s, consistently estimates the distribution *H*. Similarly, the distribution of the total time may be consistently estimated by the Kaplan-Meier estimator based on $\left(\tilde{Y}_i, \Delta_{2i}\right)$'s, and the distribution of *C* can be consistently estimated by the Kaplan-Meier estimator based on $\left(\tilde{Y}_i, 1-\Delta_{2i}\right)$'s. Because T_{23} and C_2 will be in general dependent, the estimation of the marginal distribution for the second gap time is not a simple issue. The same applies to the bivariate distribution function $F_{12}(x,y) = P(T_{12} \le x, T_{23} \le y)$.This issue have received much attention recently. Among others it was investigated by Lin et al. [5], Wang and Wells [6], Wang and Chang [7], Peña et al. [8], van der Laan et al. [9] or van Keilegom [10].

Introduce

$$\hat{F}_{12}(x,y) = \sum_{i=1}^{n} W_i I\left(\tilde{T}_{12i} \le x, \tilde{T}_{23i} \le y\right) \quad (1)$$

where $W_i = \frac{\Delta_{2i}}{n-R_i+1}\prod_{j=1}^{i-1}\left[1 - \frac{\Delta_{2j}}{n-R_j+1}\right]$ is the Kaplan-Meier weight attached to $\tilde{Y}_i$ when estimating the marginal distribution of *Y* from $\left(\tilde{Y}_i, \Delta_{2i}\right)$'s, and for which the ranks of the censored $\tilde{Y}_i$'s, R_i, are higher than those for uncensored values in the case of ties. This estimator is consistent whenever $x + y$ is smaller than the upper bound of the support of the censoring time [11]. From (1) we can obtain an estimator for the marginal distribution of the second gap time, $F_2(y) = P(T_{23} \le y)$, namely

$$\hat{F}_2(y) = \hat{F}_{12}(\infty, y) = \sum_{i=1}^{n} W_i I\left(\tilde{T}_{23i} \le y\right) \quad (2)$$

Note that estimator (2) is not the Kaplan-Meier estimator based on $\left(\tilde{T}_{23i}, \Delta_{2i}\right)$'s. This is because the weights W_i are based on the $\tilde{Y}_i$-ranks rather than on the $\tilde{T}_{2i}$-ranks. Indeed, since T_2 and C_2 are expected to be dependent, the ordinary Kaplan-Meier estimator of F_2 will be in general inconsistent.

The estimator (1) can also be written as

$$\hat{F}_{12}(x,y) = \hat{L}(x,0) - \hat{L}(x,y),$$

where

$$\hat{L}(x,y) = \frac{1}{n}\sum_{i=1}^{n} I\left(\tilde{T}_{12i} \le x, \tilde{T}_{23i} > y\right)\Delta_{2i} / \left\{1 - \hat{G}\left(\left(\tilde{Y}_i\right)^{-}\right)\right\},$$

where $\hat{G}$ is the Kaplan-Meier estimator of *G*.

The state occupation probabilities are defined as $\pi_h(t) = P(X(t) = h)$ and in particular $\pi_h(0) = P(X(0) = h)$ is the initial distribution of the process. Another quantity of interest in multi-state modelling is the cause-specific cumulative incidence function, as defined by Kalbfleisch and Prentice [12]. These quantities are appropriate if our interest is the estimation of failure probabilities. These functions are different from the transition probabilities because they represent the probability of the individual to be or having been in some particular state.

Throughout this paper we are particularly interested in the estimation of the transition probabilities. These quantities constitute one topic of much interest providing important measures to make long-term predictions.

In Markov models, the transition probabilities can be calculated from the intensities by solving the so-called forward Kolmogorov differential equation [13]. For example, for the progressive three-state model the transition probabilities $p_{hj}(s,t) = P(X(s) = j | X(s) = h)$, $h = 1,2$; $j = 1,2,3$; $h \le j$, have explicit expression,

$$p_{11}(s,t) = exp\left(-A_{12}(s,t)\right)$$
$$p_{22}(s,t) = exp\left(-A_{23}(s,t)\right)$$

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

where $A_{hj}(s,t) = \int_s^t \alpha_{hj}(u)du$ is the cumulative or integrated intensity between *s* and *t*. Note that in this model, the transition probabilities to be estimated reduce to $p_{11}(s,t)$, $p_{12}(s,t)$ and $p_{22}(s,t)$, since $p_{13}(s,t) = 1 - p_{11}(s,t) - p_{12}(s,t)$ and $p_{23}(s,t) = 1 - p_{22}(s,t)$.

In time homogeneous Markov models the explicit expressions for the transition probabilities are given by

$$p_{11}(s,t) = exp(-\alpha_{12}t)$$
$$p_{22}(s,t) = exp(-\alpha_{23}t)$$
$$p_{12}(s,t) = \frac{\alpha_{12}}{\alpha_{12}-\alpha_{23}}\left[e^{-\alpha_{23}(t-s)} - e^{-\alpha_{12}(t-s)}\right]$$

Details about the inference for the transition intensities can be seen in Andersen and Perme (2008) [4].

The transition probabilities may also be estimated via the non-parametric (Aalen–Johansen estimator) model. This can be thought as the generalization of the Kaplan-Meier approach for the simple mortality model and was proposed by Aalen and Johansen [14] for general multi-state models with a finite number of states. For the progressive three-state model, the transition probabilities $p_{11}(s,t)$ and $p_{22}(s,t)$ may be estimated by

$$\tilde{p}_{11}(s,t) = \prod_{s<t_{(k)}\le t}\left[1 - \frac{d_{12k}}{n_{1k}}\right]$$
$$\tilde{p}_{22}(s,t) = \prod_{s<t_{(k)}\le t}\left[1 - \frac{d_{23k}}{n_{2k}}\right]$$

Where $t_{(1)} < \cdots < t_{(d)}$ are the event times (which represent states 2 and 3) arranged in increased order, n_{1k} and n_{2k} denote the number of individuals in states 1 and 2, respectively, just prior to the event time $t_{(k)}$. Further, d_{12k} is the number of subjects who underwent from state 1 into state 2 at time $t_{(k)}$, while d_{23k} is the number of subjects who underwent from state 2 into state 3 at that same time.

An estimator for $p_{12}(s,t)$ is given by

$$\tilde{p}_{12}(s,t) = \sum_{s<t_{(k)}\le t}\tilde{p}_{11}(s,t_{(k-1)})\frac{d_{12k}}{n_{1k}}\tilde{p}_{22}(t_{(k)},t)$$

which is a plug-in estimator.

Datta and Satten (2001) [15] investigated the performance of the Aalen-Johansen estimator of state occupation probabilities when the process is not Markovian. These authors established the consistency of Aalen-Johansen estimators for the occupation probabilities in this case. Recently, Meira-Machado et al. (2006) [16], verified that in non-Markov situations, the use of Aalen-Johansen estimators [14] to empirically estimate the transition probabilities, $p_{hj}(s,t)$, 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 methods can easily be adapted to the progressive three-state model. Now according to the notation introduced

$$p_{11}(s,t) = P(T_{12} > t | T_{12} > s)$$
$$p_{12}(s,t) = P(T_{12} \le t, T_{12}+T_{23} > t | T_{12} > s)$$
$$p_{22}(s,t) = P(T_{12}+T_{23} > t | T_{12} \le t, T_{12}+T_{23} > s)$$

These quantities are determined by the joint distribution of (T_{12}, T_{23}). Specifically, the knowledge of the distribution of the first gap time is enough for the recovery of $p_{11}(s,t)$

$$\hat{p}_{11}(s,t) = \left(1-\hat{H}(t)\right)/\left(1-\hat{H}(s)\right) \quad (3)$$

While expectations of type $S(\phi) = E[\phi(T_{12},Y)]$ arise when handling $p_{12}(s,t)$ and $p_{22}(s,t)$:

$$\hat{p}_{12}(s,t) = \frac{1}{1-\hat{H}(s)}\sum_{i=1}^n W_i\,\phi_{s,s}\left(T_{12[i]},\tilde{Y}_i\right) \quad (4)$$

$$\hat{p}_{22}(s,t) = \sum_{i=1}^n W_i\tilde{\phi}_{s,t}\left(T_{12[i]},\tilde{Y}_i\right)/\sum_{i=1}^n W_i\tilde{\phi}_{s,s}\left(T_{12[i]},\tilde{Y}_i\right), \quad (5)$$

where W_i are the Kaplan-Meier weights attached to $\tilde{Y}_{(i)}$, with *H* the distribution function of T_{12}, and $\hat{H}$ its Kaplan-Meier estimator; $\phi_{s,t}(u,v) = I(s < u \le t, v > t)$ and $\tilde{\phi}_{s,t}(u,v) = I(u \le s, v > t)$. In these expressions, $\tilde{Y}_{(1)} \le \cdots \le \tilde{Y}_{(n)}$ denote the ordered sample of the $\tilde{Y}_i$´s, and $T_{12[i]}$ for the pair attached (concomitant) to the $\tilde{Y}_{(i)}$ value.

Estimators (4) and (5) can also be written as

$$\hat{p}_{12}(s,t) = \frac{1}{n\left(1-\hat{H}(s)\right)}\sum_{i=1}^n \frac{I\left(s < \tilde{T}_{12i} \le t, \tilde{Y}_i > t\right)\Delta_{2i}}{1-\hat{G}\left(\left(\tilde{Y}_i\right)^-\right)}$$

$$\hat{p}_{22}(s,t) = \frac{\sum_{i=1}^n \frac{I\left(\tilde{T}_{12i} \le s, \tilde{Y}_i > t\right)\Delta_{2i}}{1-\hat{G}\left(\left(\tilde{Y}_i\right)^-\right)}}{\sum_{i=1}^n \frac{I\left(\tilde{T}_{12i} \le s, \tilde{Y}_i > s\right)\Delta_{2i}}{1-\hat{G}\left(\left(\tilde{Y}_i\right)^-\right)}}$$

The estimator (4) is equivalent to Aalen-Johansen estimator.

These estimators do not rely on the Markov assumption and are motivated as natural extensions to a censored scenario of the proportion of individuals in each state. For obtaining the asymptotic results, Meira-Machado et al. (2006) [16] have used the existing theory devoted to Kaplan-Meier integrals. Special attention was paid to consistency, convergence in distribution to a normal, and (limit) variance estimation. The results can easily be adapted to the progressive three-state model. The authors evaluated this new approach through a simulation study, comparing the new method with Aalen-Johansen estimator (typically assumed in Markov situations). Results show that unless the process is, in fact, Markov, the Markov-free estimator is a wise choice. Simulations suggest that the Aalen-Johansen estimator is highly susceptible to departures from the true transition probabilities when the process is not Markov.

III. Regression models

One important goal in multi-state modelling is to study the relationships between the different predictors and the outcome. To relate the individual characteristics to the intensity rates through a covariate vector, *Z*, possibly time-dependent several models have been used in literature. The most common are: (a) Homogeneous Markov Model, (b) Piecewise Homogeneous Markov Model, (c) Cox (semi-) Markov model. Because of its simplicity, models (a) and (c) are the most used. Cox-like models (c) can be fitted through most of the statistical packages (R, S-plus, SAS, etc.), as long as a counting process notation is used, with each patient being represented by several observations [2]. The homogeneous

Markov model and the piecewise model need specialized software, most of which are written in FORTRAN, R or SAS. In this paper special attention is paid to Cox-like models. More details about the inference in models (a) and (b) can be found in Meira-Machado et al. (2008) [2].

The inference problem in a multi-state model can be decoupled into various survival models, by fitting separate intensities to all permitted transitions. For the progressive three-state model of Figure 1, assuming the process to be *Markovian*, the transition intensities, $\alpha_{hj}(t;Z)$, $h = 1,2$; $j = 1,2,3$; $h \leq j$, may be modelled using Cox-like models of the form

$$\alpha_{hj}(t;Z) = \alpha_{hj,0}(t)exp\left(\beta_{hj}^{T}Z\right) \quad (6)$$

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 is a covariate vector.

Another regression model for survival data that can easily be extended to multi-state models is the additive hazards approach of Aalen ([1], [17] and [18])

$$\alpha_{hj}(t;Z) = \alpha_{hj,0}(t) + \beta_{hj}^{T}(t)Z \quad (7)$$

where the regression functions $\beta_{hj}(t)$ are left unspecified.

By ignoring disease history behavior, Markov models such as (6) and (7) above may be inappropriate (e.g., the future health of recently diseased individuals may be different from that of individuals who have been ill for a long time). The most common deviations from the Markov property are various kinds of duration dependence. One typical approach to deal with such problem 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 [2]. Usually, this assumption is checked including covariates depending on the history. For details about other modeling approaches and assumption see for example Meira-Machado et al. (2008) [2].

Note that in multi-state models some of the transitions can be competing (for example, in the illness-death model this can happen frequently) leading to interpretational problems. This is not the case for the progressive three-state model.

The multi-state models of the form (6) and (7) are both semiparametric, and the effects of continuous predictors on log-hazard are modeled linearly. In practice, however, the effect of a given continuous predictor can be unknown, and its form may be different in all permitted transitions. If the true effect is highly nonlinear, this erroneous assumption of linearity may have serious consequences: misspecification and statistical errors (bias and decreased power of tests for statistical significance); leading to a diagnosis of nonproportional hazards. For more details see Cadarso et al. (2008) [19].

One possible approach allowing for nonlinear effects in model (6) above (and similarly to Aalen's additive model), is to express the log hazard on transition $h \rightarrow j$ as an additive Cox model of the form

$$\alpha_{hj}(t;Z) = \alpha_{hj,0}(t)exp\left(\sum f_{hj,i}(Z)\right) \quad (8)$$

where $f_{hj,i}(\cdot)$ and are smooth covariate functions. To introduce flexibility into model (8), several smoothing methods may be applied, but B-splines, smoothing splines, P-splines or Bayesian versions of P-splines are being most frequently considered in this context [20]-[22]. One particular concern in fitting these smoothing methods is the selection of reasonable values for smoothing parameters [19], [23]. In this paper, P-splines have been chosen as smoothers, since in the context of Cox-type models they behave satisfactorily when compared to other smoothers [24]. When using P-splines, automatic selection of the smoothing method based on minimizing the Akaike's information criterion [25] somewhat mitigates this problem in univariable models but cannot be used to the multivariable setting. Cadarso et al. (2008) [19] propose the use of the smoothing parameters based on a mixed model representation of penalized splines for the multivariable setting [26]. This approach, based in a Bayes empirical methodology, can be used to determine the optimal amount of smoothing in both univariable and multivariable settings. Cadarso et al. (2008) [19] propose also a flexible method for constructing smooth hazard ratio (HR) curves with confidence bands taking a specific value as the reference. For simplicity, assume a single continuous covariate. The hazard ratio for a subject with covariate value Z compared to a subject with covariate value z_{ref}, on transition $h \rightarrow j$ is given by $HR_{hj}(Z, z_{ref}) = \exp\left(f_{hj}(Z) - f_{hj}(z_{ref})\right)$. A natural estimate of the hazard ratio curve can be constructed as $\widehat{HR}_{hj}(Z, z_{ref}) = \exp\left(\hat{f}_{hj}(Z) - \hat{f}_{hj}(z_{ref})\right)$ by replacing $f_{hj}(\cdot)$ by any P-spline estimate $\hat{f}_{hj}(\cdot)$. The expression for the variance of the log hazard ratio estimate is given in terms of the covariance matrix of the smoother

$$Var\left(Ln\widehat{HR}_{hj}(Z, z_{ref})\right) =$$

$$Var\left(\hat{f}_{hj}(Z)\right) + Var\left(\hat{f}_{hj}(z_{ref})\right) - 2Cov\left(\hat{f}_{hj}(Z), \hat{f}_{hj}(z_{ref})\right)$$

Finally, assuming normality, $(1-\alpha)100\%$ pointwise confidence bands can be constructed around the hazard ratio curve

$$exp\left(Ln\widehat{HR}_{hj}(Z, z_{ref}) \pm Z_{1-\alpha/2}SE\left(Ln\widehat{HR}_{hj}(Z, z_{ref})\right)\right).$$

The use of these instruments greatly simplifies the interpretation of the continuous predictor.

IV. A REAL EXAMPLE

Due to large number of peoples affected by breast cancer, there is much demand for information on this disease. Special attention is usually paid to diagnosis [27] and prognostic forecasting [2], [19]. In a large percentage of the patients, the diagnosis is made at a sufficiently early stage when all apparent disease tissue can be surgically removed. Unfortunately, some of these patients have residual cancer, which leads to recurrence of disease and death (in some cases). Cancer patients who have experienced a recurrence are known to be at a substantially higher risk of mortality, making it essential to understand which characteristics of the patient (or of the tumour) predispose to recurrence. For the breast cancer data, we may consider the recurrence as an associated state of risk, and use the progressive three-state model with states "alive and disease-free", "alive with recurrence" and

"dead". Here, we consider a sample of subjects who underwent curative surgery for breast cancer. 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 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.

This study focussed mainly on flow cytometry parameters *DNA index* (*DNA*: ratio of the G0/G1 channel number of tumor cells to the G0/G1 channel number of diploid cells) and *S phase fraction* (*SPF*: the percentage of cells in phase S) but other factors were also included such as: age (Age: years), tumor size (*Size*: measure in mm), histologic grade (*SBR*: stages I to III), lymph node involvement (*LNI*: percentage between positive and total dissected lymph nodes), and hormone receptor status (*ER*).

A property that is often assumed when using multi-state models is the Markov property (future depends on the history only through the present). The Markov assumption may be checked by including covariates in the modelling process [28]. The results obtained for the breast cancer data show that the effect of time spent in state 1 is significant ($P < 0.05$). This allows us to conclude that Markov's model is unsatisfactory for the breast cancer. Note that the methods presented in the previous section are free of the Markov property.

In multi-state modelling, the Cox (Semi) Markov model is typically assumed whenever the interest is to study how covariates affect survival. These models can easily be fitted through most of the existing software as long as we use a counting process notation, representing each subject with several observations [2].

One main interest in breast cancer is to make diagnosis at a sufficiently early stage of the disease. Thus, is important to make long-terms predictions and to identify possible times of diagnosis (threshold values). Therefore, it is very important to obtain good estimates for the transition probabilities. Since the process for the breast cancer data does not fulfils the Markov assumption the use of the estimators proposed here is preferable. With this application, we illustrate differences between the estimated transition probability from Aalen-Johansen estimator (Markovian) and from "Markov-free" estimator. In Figure 4 we present, as an example, estimated transition probabilities for $p_{hj}(2,t)$ (above) and $p_{hj}(6,t)$ (below), $h = 1,2,\ j = 1,2,\ h \leq j$, showing that a choice between those two approaches makes a big difference (especially for the prognostic of patients with recurrence at year 6). 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 4, the "Markov-free" estimator, have fewer jump points but with 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 (from 800 days on), showing that the prognosis using the "Markov-free" estimator is poorer than Aalen-Johansen prognosis.

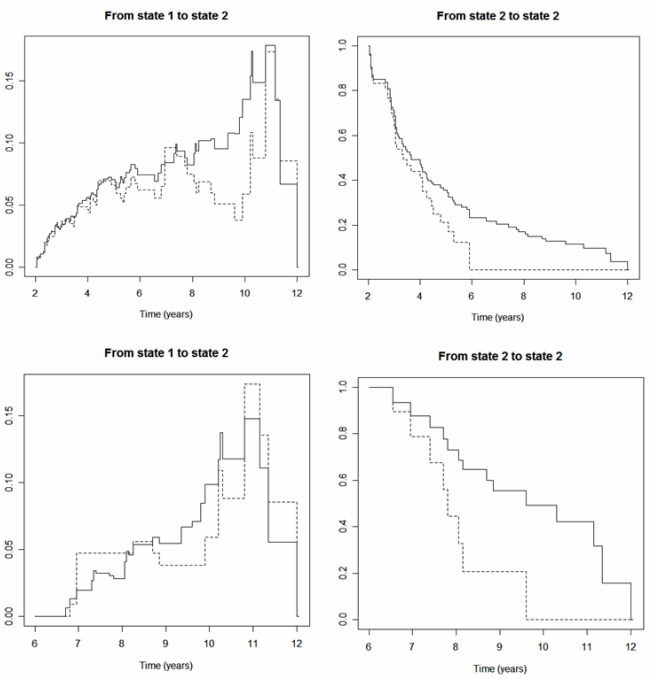

Fig. 4 Estimated transition probabilities for Aalen-Johansen estimator (solid line) and non-Markov model (dashed line).

Under gap time's framework, T_{12} is the time from randomization to cancer recurrence and T_{23} is the time from cancer recurrence to death. These times are observed subject to random right-censoring.

Table 1 presents the estimates for the joint distribution function using estimator (1) for several values of (x, y) (the x values are the percentiles 5%, 25%, 50%, 75% and maximum of the first gap time). Results show that the survival is poor for small values for the first gap time (time from randomization to recurrence). Substantially better results are obtained for higher times of recurrence.

Figure 5 illustrate the differences between the Kaplan-Meier estimator for the marginal distribution of the second gap time (based on the $(\tilde{T}_{23i}, \Delta_{2i})$'s) and estimator (2). The range of time has been restricted to 5 years to emphasize the differences between the two estimators. Differences between the two curves can be explained by the (possible) failure of the independence assumption, necessary to obtain consistency for the Kaplan-Meier estimator. Estimates for the two marginal distribution functions (using the Kaplan-Meier product-limit for the first gap time) can be used to compare the survival in two or more groups/treatments (results not shown).

Table 1: Estimates of the joint distribution function, $\hat{F}_{12}(x, y)$.

$x \setminus y$	1	2	3	4	5
0.7624	0.01852	0.02421	0.02830	0.03097	0.03097
2.9922	0.06824	0.11892	0.14127	0.16877	0.16877
4.1480	0.07538	0.13431	0.16082	0.20303	0.21068
5.6823	0.08892	0.14785	0.20356	0.24576	0.25341
12.685	0.11723	0.31946	0.38490	0.42711	0.46811

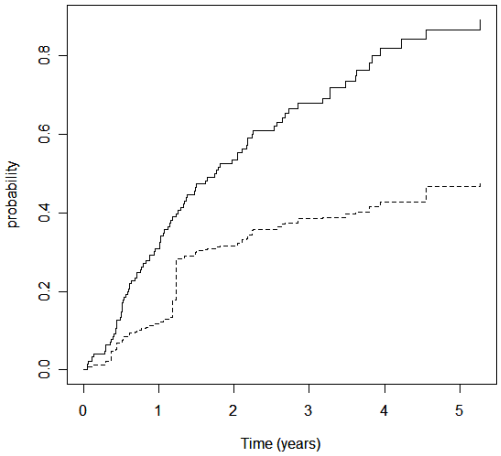

Fig. 5 Estimates of the marginal distribution function of the second gap time using our estimator (dashed line) and using Kaplan-Meier estimator (solid line).

One major goal in multi-state models is to study the relationship between the different covariates and disease evolution. These models provide more detailed information on the disease progress, by highlighting covariates affecting both mortality and recurrence. In the paper by Cadarso-Suarez et al. (2008) [19], the Galician breast cancer data is analyzed using a Cox semi-Markov model, illustrating some of the advantages of using a multi-state model. In this paper, the authors propose a flexible methodology assuming non-linear relationships between continuous predictors and survival. The use of such a multi-state model revealed important disease information regarding the effect of flow cytometry parameters considered (*DNA index* and *S phase fraction*). Comparison between the results of fitting a semi-Markov model and those obtained from the traditional Cox model yielded some important biologic insights. Among other results, this study revealed that, whereas *DNA index* is only an important predictor of recurrence intensity (similarly for *tumor size*), *S phase fraction* revealed itself a significant predictor of both recurrence and mortality. Interestingly, the effect of Age revealed to be significant only on the mortality transition (1→2). Furthermore, the important (nonlinear) effect of *DNA index* and *Tumor Size*, in the recurrence transition, would probably not have been detected by a parametric analysis (in both the linear Cox model and the linear multi-state framework). Table 2 presents the comparative results for the multivariate three models (marginal and multi-state) when introducing all the continuous predictors using P-splines (degrees of freedom using Bayes empirical methodology). More details and results (e.g., estimated effects) can be seen in Cadarso et al. (2008) [19].

Table 2: Significance of the predictors for the three Cox models (marginal Cox model; multi-state Cox semi-Markov model)

	P-vaule		
Covariate	Cox	1->2	2->3
Age (years)			
age	0.028	ns	<0.001
ps(age) nonlinear	ns	ns	ns
Size (mm)			
size	ns	ns	ns
ps(size) nonlinear	0.025	0.018	ns
LNI (%)			
LNI	<0.001	<0.001	ns
ps(LNI) nonlinear	0.037	0.019	ns
SBR			
I	-	-	-
II	ns	ns	ns
III	ns	ns	0.027
ER			
No	-	-	-
Yes	<0.001	0.004	0.049
SPF			
SPF	<0.001	<0.001	<0.001
ps(SPF) nonlinear	0.028	ns	ns
DNA			
DNA	ns	ns	ns
ps(DNA) nonlinear	0.005	<0.001	ns

ns = not significant

The presence of a nonlinear effect for *DNA* in the recurrence transition can be visually inspected and their correct functional form identified in the graphs shown in Figure 6. The plot for *DNA* suggests a spoon-shape functional form, indicating that risk decreased sharply until about 1.3, increased gradually until a value of 2.5, and then remained roughly constant.

From the biologic point of view, clinicians expect that tumors with *DNA* values close to 1 will have a better prognosis than those distant from 1. However, graph of Figure 6 shows that the main curve keeps decreasing after value 1. Although the smoothed log hazard curves shown in this figure provide important information about covariate effect on hazard, interpretation is not straightforward since we do not have a reference value. To obtain interpretable results we constructed smooth log HR curves with 95% confidence intervals to describe the relationship between the continuous predictor and risk (recurrence) for a reference value of 1. Figure 7 shows the corresponding curves showing that patients with or close to *DNA* = 1.3 had a significantly smaller risk when compared to those with *DNA* =1. One possible explanation is that a relatively large group of patients may be present with two diploid populations, with *DNA* =1.

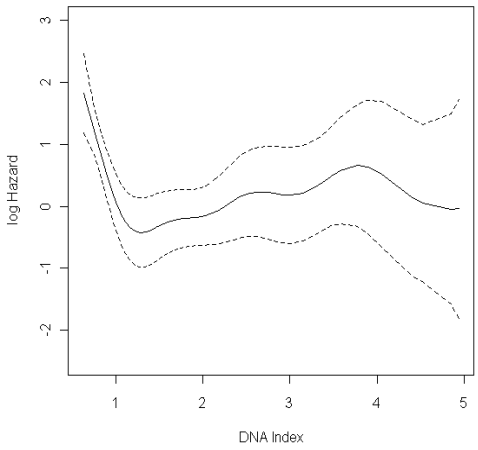

Fig. 6 Adjusted Smooth log hazard with 95% pointwise confidence bands for age, SPF and DNA Index.

The authors expressed the nonlinear relationship between continuous predictors and survival as smooth hazard ratio curves with confidence intervals where a specific value is taken as reference. The log HR curve for *DNA index* (obtained taking *DNA* = 1 as the reference value) is depicted in Figure 7. The confidence bands for the log HR are important, showing that patients with *DNA index* values lower than 1 are at higher risk, and patients with *DI* values in the interval from 1 through 1.4 are at lower risk when compared with those with *DNA* equal to 1.

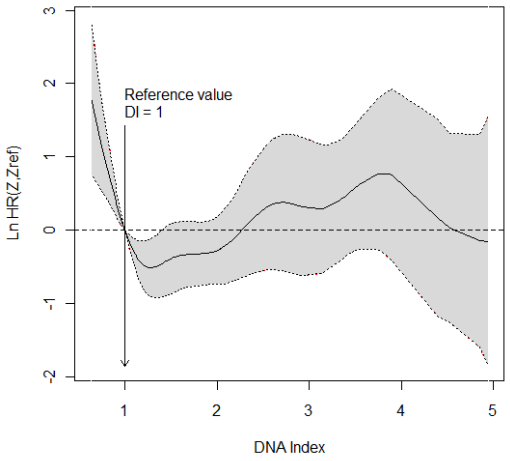

Fig. 7 Adjusted smooth log hazard ratio estimates with 95% pointwise confidence bands for DNA Index (value 1 as reference).

With regard to *Age*, a reference value of 50 years was selected as a possible value for the beginning of menopause. The corresponding plots for *Age* (see Figure 8) showed that in mortality transition the risk of death was higher for older patients, when compared with a patient aged 50 years (reference value).

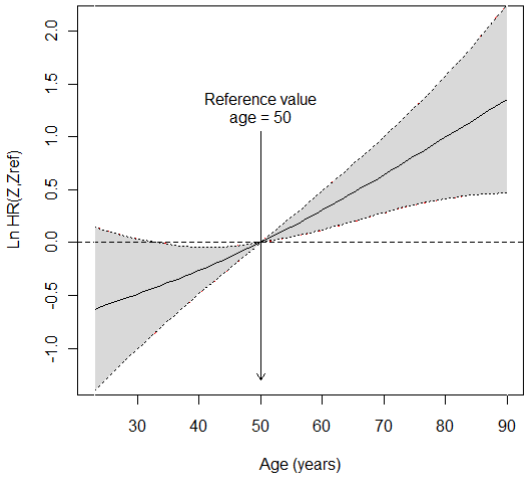

Fig. 8 Adjusted smooth log hazard ratio estimates with 95% pointwise confidence bands for Age (value 50 as reference).

V. CONCLUSION

In this paper, we have discussed the use of multi-state models in the analysis of survival data. In the scope of the progressive three-state model, nonparametric estimators are presented and illustrated using a real dataset on breast cancer from Galicia, Spain. In contrast to other existing methods, the introduced estimate for the bivariate distribution function for censored gap times (time to recurrence and time from recurrence to death) is a proper distribution function, in the sense that it attaches positive mass to each observation. We use this estimator to introduce also an estimator for the marginal distribution of the second gap time. The ideas behind the estimators are also used to introduce nonparametric estimators for the transition probabilities. The proposed methods can be easily extended for the progressive *k*-state model. In such a case we have a vector of k gap times, $(T_1, \dots, T_k)$, and the weights W_i are defined as those of the Kaplan–Meier estimator of the marginal distribution of the total time $Y = T_1 + \cdots + T_k$. The proposed method for the transition probabilities can also be adapted to more general multi-state models (e.g., illness-death, bivariate model, etc) though they can become difficult with the increase of the number of states involved. The basic ideas behind this paper can also be extended to cope with the estimation of other parameters and functions such as the cause specific cumulative incidence function.

Alternative estimators for the above quantities are given in Van Keilegom et al. (2008) [29]. This methodology assumes that the vector of gap times (T_{12}, T_{23}) satisfies the nonparametric location-scale regression model $T_{23} = m(T_{12}) + \sigma(T_{12})\varepsilon$, where the function m and σ are "smooth", and ε is independent of T_{12}. On the basis of the idea of transfer of tail information, the estimator of the error distribution is used to introduce nonparametric estimators for those targets. The asymptotic properties of these estimators are also obtained.

A flexible approach using an additive multi-state model for estimating the possible effects of continuous predictors on response is used. The application of such a model to the Galician breast cancer data illustrates the advantages of using these methods for assessing the possible effect of quantitative predictors on recurrence and mortality after recurrence. To better understand the effect of continuous covariates on the outcome results are expressed in terms of hazard ratio curves, taking a specific covariate value as reference. All analyses were performed using software (in R language) written by the authors.

ACKNOWLEDGMENT

The research was supported by Spanish Ministry of Education & Science grants MTM2005-00818 and MTM2005-01274 (European FEDER supported included) and by the Xunta de Galicia grant PGIDIT07PXIB300191PR. The research was also partially supported by CMAT and FCT under the program POCI 2010.

REFERENCES

[1] P.K. Andersen, O. Borgan , R.D. Gill, N. Keiding, *Statistical Models Based on Counting Processes*, Springer, New York, 1993.

[2] L. Meira-Machado, J. de Uña-Álvarez, C. Cadarso-Suárez , P.K. Andersen, "Multi-state models for the analysis of time-to-event data", *Statistical Methods in Medical Research* 18, to be published.

[3] P. Hougaard, "Multi-state models: a review", *Lifetime Data Analysis,* Vol. 5, 1999, pp. 239-264.

[4] P.K. Andersen, M.P. Perme, "Inference for outcome probabilities in multi-state models", *Lifetime Data Analysis*, Vol. 14, 2008, pp. 405-431.

[5] D.Y. Lin, W. Sun, Z. Ying, "Nonparametric estimation of the gap time distributions for serial events with censored data", *Biometrika,* 1986, pp.59-70.

[6] W. Wang, M.T. Wells, "Nonparametric estimation of successive duration times under dependent censoring", *Biometrika,* Vol. 85, 1998, pp. 561-572.

[7] M.C. Wang, S.H. Chang, "Nonparametric estimation of a recurrent survival function", *Journal of the American Statistical Association,* Vol. 94, 1999, pp. 146-153.

[8] E.A. Peña, R.L. Strawderman, M. Hollander, "Nonparametric estimation with recurrent event data", *Journal of the American Statistical Association,* Vol. 96, 2001, pp. 1299-1315.

[9] M.J. van der Laan, A.E. Hubbard, J.M. Robins, "Locally efficient estimation of a multivariate survival function in longitudinal studies", *Journal of the American Statistical Association,*Vol. 97, 2002, pp. 494-507.

[10] I. van Keilegom, "A note on the nonparametric estimation of the bivariate distribution under dependent censoring", *Journal of Nonparametric Statistics*, Vol. 16, 2004, pp. 659-670.

[11] J. de Unã-Álvarez, L. Meira-Machado, "A simple estimator of the bivariate distribution function for censored gap times", *Statistics and Probability Letters*, Vol.78, 2008, 2440-2445.

[12] J.D. Kalbfleisch, R.L. Prentice, *The statistical analysis of failure time data*, Wiley, New York, 1980.

[13] D.R. Cox, H.D. Miller, *The theory of stochastic processes*. London: Chapman and Hall, 1965.

[14] O. Aalen, S. Johansen, "An empirical transition matrix for nonhomogeneous Markov chains based on censored observations", *Scandinavian Journal of Statistics,* Vol. 5, 1978, pp. 141-150.

[15] S. Datta, G.A. Satten, "Validity of the Aalen-Johansen estimators of stage occupation probabilities and Nelson-Aalen integrated transition hazards for non-Markov models", *Statistics and Probability Letters*, Vol. 55, 2001, pp. 403-411.

[16] L. Meira-Machado, J. de Uña-Álvarez, C. Cadarso-Suárez, "Nonparametric estimation of transition probabilities in a non-Markov illness-death model", *Lifetime Data Analysis,* Vol. 12, 2006, pp. 325-344.

[17] O.O. Aalen, "A linear regression model for the analysis of life times", *Statistics in Medicine,* Vol. 8, 1989, pp. 907–925.

[18] D. Aydin, R.E. Omay, "A Semiparametric Additive Regression Model: Investigation of House Price in Eskisehir", *WSEAS Transactions on Mathematics*, Vol. 6, 2007, pp. 494-499.

[19] C. Cadarso-Suárez, L. Meira-Machado, T. Kneib, F. Gude, "Flexible hazard ratio curves for continuous predictors in multi-state models: a P-spline approach", *Statistical Modelling*, to be published.

[20] T.J. Hastie , R.J. Tibshirani, "Exploring the nature of covariate effects in the proportional hazards model", *Biometrics*, Vol. 46, 1990, pp. 1005-1016.

[21] P.H.C. Eilers, B.D. Marx, "Flexible smoothing with B-splines and penalties", *Statistical Science*, Vol. 11, 1996, pp. 89-121.

[22] S. Lang, A. Brezger, "Bayesian P-Splines", J*ournal of Computational and Graphical Statistics*, Vol. 13, 2004, pp. 183-212.

[23] D. Aydin, R.E. Omay, "The Smoothing Parameter Selection Problem in Smoothing Spline Regression for Different Data Sets", *WSEAS Transactions on Mathematics*, Vol. 6, 2007, pp. 477-482.

[24] U.S. Govindarajulu, D. Spiegelman, S.W. Thurston, B. Ganguli, *et al.*, "Comparing smoothing techniques in Cox models for exposure-response relationships", *Statistics in Medicine*, Vol. 26, 2007, pp. 3735-52.

[25] H. Akaike, "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, Vol. 19, 1974, pp. 716–723.

[26] T. Kneib, L. Fahrmeir, "A mixed model approach for geoadditive hazard regression", *Scandinavian Journal of Statistics*, Vol. 34, 2007, pp. 207-228.

[27] G. Khuwaja, "Breast Cancer Detection Using Mammography*"*, *WSEAS Transactions on Mathematics*, Vol. 3, 2004, pp. 317-321.

[28] R. Kay, "A Markov model for analysing cancer markers and disease states in survival studies", *Biometrics,* Vol. 42, 1986, pp. 855-865.

[29] I. Van Keilegom, J. de Uña-Álvarez, L. Meira-Machado, "Nonparametric location-scale models for successive survival times under dependent censoring", *Discussion paper* (http://www.stat.ucl.ac.be/ISpub).