Noname manuscript No.
(will be inserted by the editor)

Methods for checking the Markov condition in multi-state survival data

Gustavo Soutinho · Luís Meira-Machado

Received: date / Accepted: date

Abstract The inference in multi-state models is traditionally performed under a Markov assumption that claims that past and future of the process are independent given the present state. This assumption has an important role in the estimation of the transition probabilities. When the multi-state model is Markovian, the Aalen-Johansen estimator gives consistent estimators of the transition probabilities but this is no longer the case when the process is non-Markovian. Usually, this assumption is checked including covariates depending on the history. Since the landmark methods of the transition probabilities are free of the Markov assumption, they can also be used to introduce such tests by measuring their discrepancy to Markovian estimators. In this paper, we introduce tests for the Markov assumption and compare them with the usual approach based on the analysis of covariates depending on history through simulations. The methods are also compared with more recent and competitive approaches. Three real data examples are included for illustration of the proposed methods.

Keywords Censoring · Markov assumption · Multi-state models · Transition probabilities

1 Introduction

Multi-state models are the most suitable models for the description of complex longitudinal survival data involving several events of interest (Andersen *et al.*, 1993 [1]; Hougaard 2000 [2]; Meira-Machado *et al.*, 2009 [3]; Meira-Machado and Sestelo, 2019 [4]). A multi-state model is a model for a stochastic process, which is characterized by a finite number of states and the possible transitions among them. In general, the multi-state analysis deals

Gustavo Soutinho
Institute of Public Health of the University of Porto (ISPUP) - Rua das Taipas n.º 135, 4050-600 Porto - Portugal
Tel.: +351 222 061 820
Fax: +351 222 061 821
E-mail: gdsoutinho@gmail.com

Luís Meira-Machado
Centre of Mathematics and Department of Mathematics, University of Minho - School of Sciences - Campus de Azurém, 4800-058, Guimarães - Portugal Tel.: +351 253 510 400
Fax: +351 253 510 401
E-mail: lmachado@math.uminho.pt

with inference for transition intensities and transition probabilities. The inference for transition intensities often includes regression analysis which usually involves the modeling of each transition intensity separately. A popular choice is to model each transition intensity using a proportional hazards model assuming the process to be Markovian. This assumption claims that given the present state, the future evolution of the process is independent of the states previously visited and the transition times among them; in other words, the history of the process is summarized by the state occupied at time t. However, it has been quoted that the Markov assumption is violated in some applications (Andersen (2000, 2002) [5][6]). In such cases, if interest is on multi-state regression, one alternative approach is to use a semi-Markov model in which the future of the process does not depend on the current time but rather on the duration in the current state. Semi-Markov models are also called "clock reset" models, because each time the patient enters a new state, time is reset to 0. The Markov assumption also allows the construction of simple estimators for the transition probabilities, since individuals with different past histories become comparable (Aalen and Johansen, 1978 [7]). Unfortunately, when this assumption is violated, the use of the so-called Aalen-Johansen estimators for transition probabilities can induce bias, and thus may not be recommended. Substitute estimators for the Aalen-Johansen estimator for a non-Markov process were introduced for the first time by Meira-Machado *et al.* (2006) [8]. These authors showed that their estimators may behave more efficiently (lower mean squared errors) than the Aalen-Johansen when the Markov assumption does not hold. Allignol *et al*. (2014) [9] used a competing risks process (which is Markov) to introduce a related non-Markov estimator. Both Meira-Machado *et al*. (2006) [8] and Allignol *et al*. (2014) [9] proposals have the drawback of requiring that the support of the censoring distribution contains the support of the lifetime distribution, an assumption that is unlikely to hold in most medical applications. This line of work has been recently revisited by de Uña-Álvarez and Meira-Machado (2015) [11] who propose estimators based on subsampling, also referred to as landmarking, which are consistent regardless the Markov condition and the referred assumption on the censoring support. Putter and Spitoni (2018) [12] recover the work by de Uña-Álvarez and Meira-Machado (2015) [11] to propose alternative non-Markovian estimation methods which are based on the landmark methodology combined with the Aalen-Johansen estimate of the state occupation probabilities derived from the same subsamples. The ideas of subsampling were also used by Titman (2015) [10] who extended and improved the estimator proposed by Allignol *et al*. (2014) [9].

To perform inference for transition intensities or for the transition probabilities it is essential to check if the Markov assumption is tenable. This assumption is usually checked by including covariates depending on the history (Kay, 1986 [13]; Andersen *et al.* 2000, 2002 [5][6]). For the progressive illness-death model, for example, the Markov assumption is particularly relevant for modeling death transition after disease and consequently to assess whether this transition rate is affected by the time in the previous state. Alternative methods, based on a local Kendall's tau, measuring the future-past association along time, were proposed by Rodríguez-Girondo and de Uña-Álvarez (2012, 2016) [14][15]. These methods can be used for three-state progressive and illness-death models but the extension of these tests to general multi-state models is not straightforward and thus, flexible methods that may be used in general models are required. A very recent work by Titman and Putter (2020) [16] considers new approaches to check this assumption. In one of these approaches a general test is developed by considering summaries from families of log-rank statistics where patients are grouped by the state occupied at different times. Chiou et al. (2018) [17] also considered an equivalent problem for testing Markoviaty (in the progressive illness-model) but involving tests for dependent truncation.

The organization of this paper is as follows. The following section provides an introduction to the methodological background and introduces tests for checking the markov assumption. In Section 3, we evaluate the performance of the proposed methods and compare them with competive methods through simulations studies. In Section 4, the use of the proposed methods is illustrated by the analysis of an illness-death model describing the disease process of breast and colon cancer patients. Liver cirrhosis data is used to illustrate the application of the proposed methods to more general models. Main conclusions and discussion are reported in Section 5.

2 Multi-state models

A multi-state model is a model for a time continuous stochastic process $(X(t), t \in [0, \infty))$, taking values in the state space $\mathcal{S} = 1, \ldots, K$, with K finite, and fulfilling some simplification assumptions. This multi-state process is fully characterized through transition probabilities between states h and j, that we express by $p_{hj}(s,t|\mathcal{H}_{s-}) = P(X(t) = j|X(s) = h, \mathcal{H}_{s-})$, for $h, j \in \mathcal{S}$ and $s < t$, where $\mathcal{H}_{s-}$ denotes the history of the multi-state process up to s. In particular, the history of the 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 is also characterized through the transition intensities $\lambda_{hj}(t|\mathcal{H}_{t-}) = \lim_{\Delta t \to 0} P(X(t+\Delta t) = j|X(t) = h, \mathcal{H}_{t-})/\Delta t$ which can be considered as a generalization of the hazard function in survival analysis. The cumulative transition intensities are defined as $\Lambda_{hj}(t) = \int_0^t \lambda_{hj}(u)du$, with $\Lambda_{hh}(t) = -\sum_{j \neq h} \Lambda_{hj}(t)$ the (h,h)th diagonal element of the $K \times K$ matrix $\boldsymbol{\Lambda}(t)$. Similarly, define the $K \times K$ matrix $\mathbf{P}(s,t)$ with the (h,j)th element $p_{hj}(s,t)$.

When the multi-state process is Markov, the transition intensities simplifies to $\lambda_{hj}(t) = \lim_{\Delta t \to 0} P(X(t+\Delta t) = j|X(t) = h)/\Delta t$ and the transition probabilities to $p_{hj}(s,t) = P(X(t) = j|X(s) = h)$. In particular this means that under the Markov assumption, $P(X(t) = j|X(s) = h, X(u) = y) = P(X(t) = j|X(s) = h)$ for any $0 \leq u < s$ and $y \in \mathcal{S}$, and thus, that the future of the process after time s depends only on the state occupied at time s, not on the arrival time to that state or on the states previously visited.

For Markovian processes, the transition probability matrix $\mathbf{P}(s,t)$ can be recovered from the transition intensities through product integration (Aalen and Johansen, 1978 [7]):

$$\mathbf{P}(s,t) = \prod_{s<u\leq t} (\mathbf{I} + d\boldsymbol{\Lambda}(u))$$

where $\mathbf{I}$ is the $K \times K$ identity matrix, and where the cumulative transition intensities can be estimated by the Nelson-Aalen estimator (Andersen *et al.*, 1993 [1])

$$\widehat{\Lambda}_{hj}(t) = \int_0^t \frac{N_{hj}(t)}{Y_h(t)},$$

where $N_{hj}(t)$ is the number of observed direct transitions from state h to state j up to time t and $Y_h(t)$ is the number of individuals under observation in State h just before time t. Then, the Aalen-Johansen estimator takes the form

$$\widehat{\boldsymbol{P}}(s,t) = \prod_{s<u\leq t} \left(\mathbf{I} + d\widehat{\boldsymbol{\Lambda}}(u)\right)$$

For simple models like the illness-death model, we can give explicit expressions for the elements of $\widehat{\boldsymbol{P}}(s,t)$. Expressions for general models are not possible.

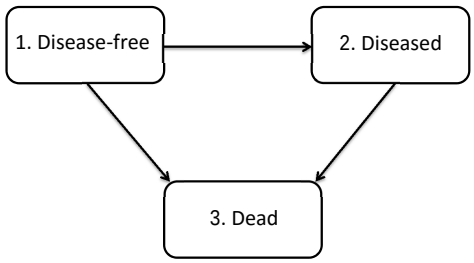

Fig. 1 The progressive illness-death model.

2.1 The progressive illness-death model

Without loss of generality and for the purpose of simplicity, from this point on, we will consider the progressive illness-death model depicted in Figure 1. This model is encountered in many medical studies for describing the progression of patients undergoing a given illness, particularly in cancer studies such as for our two real data examples. Many time-to-event data sets from medical studies with multiple end points can also be reduced to this generic structure. In this model, individuals are at risk of death in each transient state (states 1 and 2). The illness-death model is not necessarily Markovian, since the prognosis for an individual in the intermediate state may be influenced by the subject specific arrival time.

The progressive illness-death model is characterized by three transition intensities: the disease intensity $\lambda_{12}(t)$, the mortality intensity without the disease $\lambda_{13}(t)$ and the mortality intensity among the diseased individuals, $\lambda_{23}(t, t_{12})$. The later transition intensity may depend on t_{12}, the time of the disease occurrence in the illness-death process: $\lambda_{23}(t, t_{12}) = \lim_{\Delta t \to 0} P(X(t + \Delta t) = 3 | X(t) = 2, T_{12} = t_{12})/\Delta t$ where T_{12} represent the potential transition times from State 1 to State 2. The process is called Markov if $\lambda_{23}(t, t_{12})$ is independent of t_{12}, otherwise it is called semi-Markov (i.e., future evolution not only depends on the current state, but also on the entry time into that same state).

In the particular case of the progressive illness-death model the transition probabilities can be obtained from the transition intensities as follows (Beyersmann, Schumacher and Allignol, 2012)[18]

$$p_{11}(s,t) = \exp\left(-\int_s^t (\lambda_{12}(u) + \lambda_{13}(u))\, du\right)$$

$$p_{22}(s,t \mid t_{12}) = \exp\left(-\int_s^t \lambda_{23}(u, t_{12}) du\right)$$

$$p_{12}(s,t) = \int_s^t p_{11}(s, u-)\lambda_{12}(u) p_{22}(u, t \mid u) du.$$

Here, $p_{22}(s, t \mid t_{12})$ denotes the transition probability p_{22} conditionally on a particular entry time t_{12}. If the process is Markov, $h_{23}(t, t_{12}) = h_{23}(t)$ and $p_{22}(s, t \mid t_{12}) = p_{22}(s, t)$. The two other transition probabilities $p_{13}(s, t)$ and $p_{23}(s, t)$ can be estimated from the two obvious relations that exist in the progressive illness-death model: $p_{11}(s, t) + p_{12}(s, t) + p_{13}(s, t) = 1$ and $p_{22}(s, t) + p_{23}(s, t) = 1$.

The Aalen-Johansen estimator is the standard nonparametric estimator of the transition probabilities for Markov processes. Explicit formulae of the Aalen-Johansen estimator (Aalen and Johansen (1978) [7]) for the illness-death model are given are given by the following expressions:

$$\widehat{p}_{11}^{\,\mathrm{AJ}}(s,t) = \prod_{s<t_i\leq t}\Big(1 - dN_1(t_i)/Y_1(t_i)\Big),$$

$$\widehat{p}_{22}^{\,\mathrm{AJ}}(s,t) = \prod_{s<t_i\leq t}\Big(1 - dN_{23}(t_i)/Y_2(t_i)\Big),$$

and

$$\widehat{p}_{12}^{\,\mathrm{AJ}}(s,t) = \sum_{s<t_i\leq t}\widehat{p}_{11}^{\,\mathrm{AJ}}(s,t_i^-)\frac{dN_{12}(t_i)}{Y_1(t_i)}\widehat{p}_{22}^{\,\mathrm{AJ}}(t_i,t)$$

Where $dN_1(t_i) = dN_{12}(t_i) + dN_{13}(t_i)$ for the total number of transitions out of state 0 and let $Y_1(t_i)$ and $Y_2(t_i)$ be the number of healthy (i.e. in state 1) and diseased (i.e. in state 2) individuals, respectively, just prior to time t_i. Since $\widehat{p}_{11}^{\,\mathrm{AJ}}(s,t)$ and $\widehat{p}_{22}^{\,\mathrm{AJ}}(s,t)$ are Kaplan-Meier estimators, their variance may be estimated by Greenwood's formula. The expression for the variance of $\widehat{p}_{12}^{\,\mathrm{AJ}}(s,t)$ can be found in (Borgan, 2005) [19].

When the multi-state model is Markovian, the Aalen-Johansen estimator gives consistent estimators of the transition probabilities. This estimator may also be used to consistently estimate occupation probabilities for non-Markov multi-state models (Datta and Satten, 2001 [20]). When the process is not Markovian, the Aalen-Johansen estimator of the transition probabilities may introduce some bias and therefore they may be inappropriate. Estimators that do not rely on the Markov assumption were first introduced by Meira-Machado et al. (2006) [8]. The proposed estimators were defined in terms of multivariate Kaplan-Meier integrals, and proven to be more efficient than the Aalen-Johansen estimators in case of strong violation of the Markov assumption. Since then, there has been several contributions on the topic but two recent papers stand out. The first paper, by de Uña-Álvarez and Meira-Machado (2015) [11], uses the ideia of subsampling, also referred to as landmarking (Van Houwelingen, 2007) [21], which is based on (differences between) Kaplan-Meier estimators derived from a subset of the data consisting of all subjects observed to be in the given state at the given time. To be specific, in the illness-death model, given the time point s, to estimate $p_{1j}(s,t)$ for $j = 1,2,3$ the landmark analysis is restricted to the individuals observed in State 1 at time s; whereas, to estimate $p_{2j}(s,t)$, $j = 2,3$, the landmark analysis proceeds from the sample restricted to the individuals observed in State 2 at time s. The non-Markov illness-death model is characterized by the joint distribution of (Z,T), where Z is the sojourn time in the initial state 1 and T is the total survival time. Under censoring, only the censored versions of Z and T, along with their corresponding censoring indicators, are available. Define $\widetilde{Z} = \min(Z,C)$ and $\widetilde{T} = \min(T,C)$, where C is the potential censoring time, which is assumed to be independent of (Z,T). For the illness-death model, we may then formally introduce the landmark estimators introduced by de Uña-Álvarez and Meira-Machado (2015) [11] as follows

$$\widehat{p}_{11}^{\,\mathrm{LM}}(s,t) = \widehat{S}_0^{\mathrm{KM}(s)}(t),\ \widehat{p}_{22}^{\,\mathrm{LM}}(s,t) = \widehat{S}^{\mathrm{KM}[s]}(t)$$

$$\widehat{p}_{12}^{\,\mathrm{LM}}(s,t) = \widehat{S}^{\mathrm{KM}(s)}(t) - \widehat{S}_0^{\mathrm{KM}(s)}(t),\quad \widehat{p}_{23}^{\,\mathrm{LM}}(s,t) = 1 - \widehat{S}^{\mathrm{KM}[s]}(t)$$

$$\widehat{p}_{13}^{\,\mathrm{LM}}(s,t) = 1 - \widehat{S}^{\mathrm{KM}(s)}(t)$$

where $\widehat{S}_0^{\mathrm{KM}(s)}$ and $\widehat{S}^{\mathrm{KM}(s)}$ are the Kaplan-Meier estimators for the distributions of Z and T, respectively, but computed from the subsample $\mathcal{S}_1 = \Big\{i : \widetilde{Z}_i > s\Big\}$; whereas $\widehat{S}^{\mathrm{KM}[s]}$

is the Kaplan-Meier estimator of the distribution of T but computed from the subsample $\mathcal{S}_2 = \left\{ i : \widetilde{Z}_i \leq s < \widetilde{T}_i \right\}$.

The subsampling approach combined with the Aalen-Johansen estimate of the state occupation probabilities was later used by Putter and Spitoni (2018) [12] to introduce the termed Landmark Aalen-Johansen estimator. The landmark Aalen-Johansen estimators of the transition probabilities may then be introduced as

$$p_{hj}^{\texttt{LMAJ}}(s,t) = \widehat{\pi}^{\texttt{LM}}(s) \prod_{s<u\leq t} \left(\mathbf{I} + d\widehat{\boldsymbol{\Lambda}}^{\texttt{LM}}(u) \right)$$

with $\widehat{\pi}^{\texttt{LM}}(s)$ a $1 \times K$ vector with $\widehat{\pi}^{\texttt{LM}}(s) = 1$ for the jth element, and other values equal to 0. Here, the estimator of the cumulative transition intensities, $\widehat{\Lambda}^{\texttt{LM}}$, is Nelson-Aalen estimator computed on a landmark data set which selects subjects observed to be in State h at time s (Putter and Spitoni, 2018) [12]. Though the two landmark methods (abbreviated by `LM` and `LMAJ`) do not rely on the Markov condition, they usually lead to estimators with higher variability. Since the Aalen-Johansen reports a smaller variance in estimation, this approach should be preferred over non-Markovian estimators when one is confident of the Markov assumption.

From now on we will use the abbreviation `AJ` for the Aalen- Johansen estimator, `LM` for the LandMark estimator proposed by de Uña-Álvarez and Meira-Machado (2015) [11], and `LMAJ` for the LandMark Aalen-Johansen estimator proposed by Putter and Spitoni (2018) [12].

It is worth mentioning that, in the progressive illness-death model, the results of the three estimators of the transition probability $p_{11}(s,t)$ are equal. Minor differences are appreciated when comparing the `LM` and `LMAJ` estimators for the remaining transition probabilities.

2.2 Tests for the Markov assumption

Traditionally, the Markov condition is verified by modeling particular transition intensities on aspects of the history of the process using a proportional hazard model (Kay, 1986) [13]. In the progressive illness-death model, for example, we can examine whether the time spent in the initial state is important on the transition from the disease state (the intermediate state) to death (the absorbing state) or not. For doing that, let $\lambda_{23}(t)$ denote the hazard function of T for those individuals going from State 2 to State 3, and let Z denote the sojourn time in State 1. Fitting a Cox model $\lambda_{23}(t \mid Z) = \lambda_{23,0}(t)\exp(\beta Z)$, where $\lambda_{23,0}$ is the baseline hazard and β a regression parameter, we now need to test the null hypothesis, $H_0 : \beta = 0$, against the general alternative, $H_1 : \beta \neq 0$. This would assess if the transition rate from the disease state into death is unaffected by the time spent in the initial state. It is worth to remember that the semiparametric Cox proportional hazard model is based on the assumption of proportional hazards and that it assumes a linear effect on the hazard for the covariate. Both may fail in practice, and consequently this approach may be unable to detect the lack of Markovianity.

Since the landmark methods (`LM`) for estimating the transition probabilities proposed by de Uña-Álvarez and Meira-Machado (2015) [11], and (`LMAJ`) by Putter and Spitoni (2018) [12] are free of the Markov assumption, they can also be used to introduce local tests for Markovianity by measuring their discrepancy to Markovian Aalen-Johansen estimators (AJ), for a fixed value $s > 0$. Though the two landmark methods behave similarly, the `LMAJ` can be used in general multi-state models which can be considered an advantage.

These ideas were recently used by Titman and Putter (2020) [16] to introduce tests based on summaries from families of log-rank statistics where patients are grouped by the state occupied at a given (landmark) time.

In this paper we also introduce a local test based on the areas under the two curves, AUC, (i.e., the curves of the estimated transition probabilities) that can be used for a general multi-state model. We propose the use of the following test statistic based on direct nonparametric estimates of the transition probabilities, $U = \int_s^\tau \left(\widehat{p}_{hj}^{\,\texttt{LMAJ}}(s,u) - \widehat{p}_{hj}^{\,\texttt{AJ}}(s,u) \right) du$, where τ is the upper bound of the support of T. The test statistic can be seen as the difference between the areas under the estimated transition probability curve for the non-markov `LMAJ` estimator and the `AJ` estimator. Intuitively, the test statistics should be close to zero if the process is Markov. The Markov assumption becomes less likely as the test statistic get further away from zero in either direction. Because of censoring, both estimators (`LMAJ` and `AJ`) may reveal high variability in the right tail which may inflate the test statistic. In addition to this issue, since landmarking is based on reduced data, the maximum point for which the `LMAJ` transition probability estimate is strictly defined may be lower than the maximum point for `AJ`. To overcome these problems, we suggest that in the computation of U one should use the minimum between the upper bound for which `LMAJ` is defined and the 90% percentile of the total time for the upper limit in the integral that defines the test statistic. In the progressive illness-death model, besides the transition probability $\widehat{p}_{23}(s,t)$, also $\widehat{p}_{12}(s,t)$ can be used to test the Markov assumption. For general multi-state models, one can use transitions depending on history (i.e., $p_{hj}(s,t)$ depending on subject specific arrival time at state $h > 1$). In fact, if the goal is to decide which estimator is the most appropriate to use to estimate a specific transition probability $p_{hj}(s,t)$, then the test statistic should be the one based on that same transition probability.

Note that if the null hypothesis of Markovianity holds, the value of U should be close to zero. To approximate the distributions of the test statistic, bootstrap methods with a large number of resamples, M, are used. We generate M bootstrap samples and for each sample calculate the test statistic $U^\star$. Then, according to large sample asymptotic distribution theory, when M, the number of replicates goes to infinity, we have the following statistic distributed approximately as a standard normal distribution with a mean of 0 and variance of 1: $V = (U-0)/\sigma^\star_{(U^\star)} \sim N(0,1)$. The null hypothesis will be rejected if $V > v_{(1-\alpha/2)}$ or $V < v_{(\alpha/2)}$, where $v_{(\alpha/2)}$ and $v_{(1-\alpha/2)}$ denote the $\alpha/2$ and $1-\alpha/2$ percentiles, respectively, of a normal distribution with a mean of 0 and variance of 1.

In this paper we also propose a global test which can be achieved by combining the results obtained from local tests over different times. The testing procedure used here involves the following steps:

Step 1: Using the original sample of the illness-death model, obtain the percentiles 5, 10, 20, 30 and 40 of the sojourn time in State 1. For general multi-state models, we recommend the use of the same percentiles of the subject specific arrival time at the corresponding state.

Step 2: For each of the values s obtained in Step 1, obtain the probability values for the local method as explained before.

Step 3: Obtain the mean of the probability values for each closest pairs; i.e., the mean of the probability values of the following pairs of percentiles: $(5,10)$, $(10,20)$, $(20,30)$ and $(30,40)$.

Step 4: Get the minimum between the four probability values obtained in Step 3.

Step 1 considers a global test based on local tests computed at low percentiles of subject specific arrival times at the corresponding state. This is based on our experience that the failure of Markovianity often occurs for small transition times. Besides the hypothesis tests

proposed above, in Section 4 we also propose graphical local tests that can be used to check the Markov assumption in the illness-death model as well as for more complex multi-state models, possibly with reversible transition between states. These graphical tests can be used to validate the default values proposed in Step 1 or to propose alternative values for which a discrepancy between the two methods (`LMAJ` and `AJ`) is more evident. The procedure described in Step 3 can be used to ensure that there is a discrepancy between the two estimated curves in a large range of time values.

To provide the biomedical researchers with an easy-to-use tool to compute the proposed methods, we are currently developing an R package which will be available at the CRAN repository. The package will allow users to choose different percentiles for the sojourn time in State 1 (Step 1). A preliminary version of this library will be provided upon request.

3 Simulation study

In this section we report results from simulation studies, where the aim is to compare the finite sample performance of the proposed methods to test the Markov assumption in a progressive illness-death model. Due to computing time issues the simulations shown here only address this model. However, an application of the proposed methods to a more complex multi-state model is presented in Section 4 from a real data set. To simulate the data in the progressive illness-death model, we assume that all individuals are in the initial state (State 1) at time $t = 0$, and that these individuals may follow two possible paths: passing through the intermediate state (State 2), at some specific time; or going directly to the absorbing state (State 3). Transition times for those leaving the initial state are generated from the cause-specific hazards given by $\lambda_{12}(t) = 0.29/(t+1)$ and $\lambda_{13}(t) = 0.024 \times t$, where $t > 0$ denotes the time since the start point. To study the Markov assumption, three different scenarios are considered corresponding to different hazards that are used to generate death times for individuals passing through the intermediate state: $\lambda^1_{23}(t) = 0.05$, $\lambda^2_{23}(t) = 0.25 \times (t_{12}+1)^{-0.8}$ and $\lambda^3_{23}(t) = 0.04 \times \log(t+1)$, where t_{12} is the transition time to the intermediate event. The first scenario is Markov since the hazard is independent of time, whereas the second is semi-Markov and the third is non-Markov. Censoring times were generated from uniform distributions. Two samples size were considered for each scenario ($n = 250$ and $n = 500$).

We also consider a fourth scenario in which the traditional test, based on the Cox proportional hazard model may fail. In this scenario, the transition times are generated from the following cause-specific hazards given by $\lambda_{12}(t) = 1/(2-t)$, $\lambda_{13}(t) = 2/(3-2t)$ for $0 \leq t < 2$ and $0 \leq t < 1.5$, respectively. To generate death times for individuals passing through the intermediate state we consider $\lambda_{23}(t) = \exp\left(-(t_{12}-1)^2\right)$. This simulated scenario is the same as that described in Rodríguez-Girondo and Uña-Álvarez (2016)[15]. Note that this scenario is non-Markov because of the dependence on the transition time to the intermediate state but in this case a misspecification of the Cox model is expected because of the shape of the hazard $\lambda_{23}(t)$ with a parabolic influence of the predictor.

Table 1 reports the rejection proportions of the proposed tests for the first three scenarios with sample sizes $n = 250$ and $n = 500$. Random censoring was simulated using uniform distributions $U[0, 60]$ and $U[0, 30]$. The first censoring distribution led to medium censoring percentages (between 41% and 47%) whereas these percentages increase in the second censoring distribution (between 45% and 62%). Four tests are considered in this table: (i) local test based on the area under the transition probabilities $\widehat{p}_{12}(s,t)$ and $\widehat{p}_{23}(s,t)$, denoted by AUC(s); (ii) local test proposed by Titman and Putter (2020) [16], based on the log-rank,

Table 1 Rejection proportions for nominal level of 5% of the local tests for fixed values $s = 1$, $s = 2$, $s = 4$, $s = 6$ and $s = 8$ (AUC(s) and LR(s)). Rejection proportions for the global tests (AUC and Cox) are also included. Censoring times uniformly distributed between 0 and 30, and between 0 and 60.

										Global	
Scenario	Trans. Prob.	n	C	Method	1	2	4	6	8	AUC / LR	Cox
Markov	$\widehat{p}_{12}(s,t)$	250	$U[0,30]$	AUC(s)	0.055	0.055	0.064	0.073	0.062	0.073	0.046
		500	$U[0,30]$	AUC(s)	0.066	0.057	0.069	0.072	0.076	0.057	0.045
	$\widehat{p}_{23}(s,t)$	250	$U[0,30]$	LR(s)	0.051	0.047	0.054	0.051	0.056	0.043	0.046
		500	$U[0,30]$	LR(s)	0.036	0.048	0.054	0.052	0.057	0.052	0.045
	$\widehat{p}_{23}(s,t)$	250	$U[0,30]$	AUC(s)	0.055	0.043	0.049	0.046	0.033	0.076	0.046
		500	$U[0,30]$	AUC(s)	0.060	0.052	0.061	0.065	0.055	0.056	0.045
Semi-Markov	$\widehat{p}_{12}(s,t)$	250	$U[0,30]$	AUC(s)	0.765	0.762	0.611	0.437	0.286	0.845	0.757
		500	$U[0,30]$	AUC(s)	0.964	0.961	0.881	0.701	0.530	0.992	0.977
	$\widehat{p}_{23}(s,t)$	250	$U[0,30]$	LR(s)	0.872	0.891	0.739	0.520	0.296	0.960	0.757
		500	$U[0,30]$	LR(s)	0.996	0.999	0.976	0.862	0.635	1.000	0.977
	$\widehat{p}_{23}(s,t)$	250	$U[0,30]$	AUC(s)	0.759	0.744	0.536	0.316	0.131	0.862	0.757
		500	$U[0,30]$	AUC(s)	0.967	0.955	0.855	0.648	0.449	0.993	0.977
non-Markov	$\widehat{p}_{12}(s,t)$	250	$U[0,30]$	AUC(s)	0.172	0.284	0.308	0.292	0.258	0.354	0.382
		500	$U[0,30]$	AUC(s)	0.336	0.458	0.508	0.502	0.468	0.602	0.701
	$\widehat{p}_{23}(s,t)$	250	$U[0,30]$	LR(s)	0.225	0.268	0.267	0.241	0.191	0.414	0.382
		500	$U[0,30]$	LR(s)	0.369	0.464	0.515	0.479	0.384	0.696	0.701
	$\widehat{p}_{23}(s,t)$	250	$U[0,30]$	AUC(s)	0.172	0.240	0.226	0.176	0.114	0.302	0.382
		500	$U[0,30]$	AUC(s)	0.348	0.452	0.474	0.420	0.332	0.574	0.701
Markov	$\widehat{p}_{12}(s,t)$	250	$U[0,60]$	AUC(s)	0.048	0.038	0.048	0.050	0.072	0.066	0.058
		500	$U[0,60]$	AUC(s)	0.052	0.052	0.050	0.042	0.070	0.062	0.038
	$\widehat{p}_{23}(s,t)$	250	$U[0,60]$	LR(s)	0.055	0.055	0.061	0.053	0.054	0.043	0.058
		500	$U[0,60]$	LR(s)	0.064	0.067	0.053	0.054	0.052	0.046	0.038
	$\widehat{p}_{23}(s,t)$	250	$U[0,60]$	AUC(s)	0.048	0.036	0.042	0.044	0.068	0.062	0.058
		500	$U[0,60]$	AUC(s)	0.050	0.054	0.050	0.032	0.068	0.062	0.038
Semi-Markov	$\widehat{p}_{12}(s,t)$	250	$U[0,60]$	AUC(s)	0.918	0.946	0.84	0.736	0.600	0.980	0.926
		500	$U[0,60]$	AUC(s)	0.998	1.000	0.982	0.940	0.876	1.000	0.940
	$\widehat{p}_{23}(s,t)$	250	$U[0,60]$	LR(s)	0.961	0.970	0.943	0.847	0.708	0.998	0.926
		500	$U[0,60]$	LR(s)	1.000	1.000	0.999	0.999	0.961	1.000	0.940
	$\widehat{p}_{23}(s,t)$	250	$U[0,60]$	AUC(s)	0.918	0.928	0.812	0.664	0.490	0.982	0.926
		500	$U[0,60]$	AUC(s)	0.996	1.000	0.982	0.942	0.848	1.000	0.940
non-Markov	$\widehat{p}_{12}(s,t)$	250	$U[0,60]$	AUC(s)	0.282	0.382	0.442	0.410	0.382	0.504	0.368
		500	$U[0,60]$	AUC(s)	0.474	0.652	0.724	0.724	0.656	0.754	0.692
	$\widehat{p}_{23}(s,t)$	250	$U[0,60]$	LR(s)	0.260	0.341	0.431	0.412	0.376	0.546	0.368
		500	$U[0,60]$	LR(s)	0.504	0.650	0.739	0.711	0.663	0.861	0.692
	$\widehat{p}_{23}(s,t)$	250	$U[0,60]$	AUC(s)	0.276	0.344	0.404	0.330	0.288	0.472	0.368
		500	$U[0,60]$	AUC(s)	0.506	0.648	0.692	0.704	0.656	0.758	0.692

for the transition probability $\widehat{p}_{23}(s,t)$, denoted by LR(s); (iii) global test based on the area under the transition probabilities (AUC) or the global test base on the log-rank statistics (LR) (Titman and Putter (2020) [16]); (iv) global test based on the Cox model (Cox). The global test LR is based on the mean value of the log-rank statistics as described in Titman and Putter (2020) [16]. The local tests were evaluated at five fixed values $s = 1$, $s = 2$, $s = 4$, $s = 6$ and $s = 8$. Results in this table were obtained by the empirical rejection proportions from 1000 trials at the significant level of 0.05.

Results show that, for the semi-Markov and non-Markov scenarios, the power of the tests is higher for lower censoring percentages, increasing with the sample size. The bootstrap test based on the areas under the curves (of the transition probabilities) (AUC) and the local test based on log-rank statistics both reveal their capacity to identify the differences between curves in the semi-Markov scenario showing higher rejection probabilities for lower values of s. Note that in this scenario, departures between the two curves (obtained from `AJ` and `LMAJ` methods) are expected to decrease as the difference $t-s$ increase. In non-Markov scenario, departures between the two curves (obtained for the transition probabilities $\widehat{p}_{12}(s,t)$ and $\widehat{p}_{23}(s,t)$ from `AJ` and `LMAJ` methods) denote a great improvement when considering a

sample size of n = 500, but with rejection probabilities below 0.50 for all s, with the exception for censoring uniform distribution $U[0, 60]$. Both local tests also obtain low rejection proportions (near the nominal level of 5%) when the data is generated from a Markov scenario. Note that we expect rejection proportions about 0.05 in this case. The results based on the log-rank statistic also confirm the good accuracy of this method in agreement with the conclusions shown in Titman and Putter (2020) [16]. In general for all scenarios, sample sizes and censoring distributions, results between the log-rank test and the local AUC test are quite similar being able to distinguish the inequality between `AJ` and `LMAJ` curves in semi-Markov and non-Markov scenarios, while providing low rejection proportions when the process is indeed Markovian. When comparing the results for the two local tests based on different transition probabilities, $\widehat{p}_{12}(s,t)$ and $\widehat{p}_{23}(s,t)$, it can be seen that they provide similar values but slightly higher when based on the computation of the transition probability $\widehat{p}_{12}(s,t)$. This behavior may be explained by the number of observations from which the transition probability is computed, those in State 1 at time s for $\widehat{p}_{12}(s,t)$, and those in State 2 at the same time for $\widehat{p}_{23}(s,t)$. For completeness purposes, Table 1 also show the results from the three global tests. These global tests present satisfactory results in all scenarios, reporting rejection proportions of about 5% for the Markov scenario, and high levels of rejection proportions for the semi-Markov and non-Markov scenarios. These results are in accordance with those obtained using a local test based on the area under the curves of the estimated transition probabilities. As expected, in general, the performance of the proposed methods is improved for scenarios with less censoring percentages (i.e., for censoring times following an uniform distribution $U[0, 60]$). This improvement is not so obvious for the method based in the Cox model. We can also notice that the global log-rank and the AUC global tests behave similarly in all cases. Some of these patterns, for censoring uniform distribution $U[0, 30]$, can be clearly seen in Figure 2.

Table 2 reports the rejection proportions of the four proposed tests for the fourth scenario, non-Markovian with an hazard with a quadratic predictor. Random censoring was simulated from uniform distributions $U[0, \tau_G]$ for τ_G equal to 8.1 and 4.6. The model with $\tau_G = 8.1$ results in 12% censoring on the first gap time and in 24% for the total time. The model with $\tau_G = 4.6$ increases these censoring levels to 20% and about 40%, respectively. In this case, the global method based on the Cox proportional model has a bad performance which can be explained by failure of the linear specification of the Cox model. It can also be seen that the power of this test does not increase substantially with the sample size, as it happens in semi-Markov and non-Markov scenarios shown in Table 1. Results shown in Table 2, reveal that the tests (local and global) based on the area under the curves have a good performance, revealing reasonable levels of rejection proportions of Markovianity. It can be seen that the power of these tests increase with the sample size. Results in terms of power performance for non-Markovian scenario, hazard with a quadratic predictor are shown in Figure 3. The plots show the rejection probabilities for the transition probability $\hat{p}_{23}(s,t)$ as a function of s. Simulation results also confirm the similarity of the local and global tests between the log-rank and the AUC test for both scenarios.

Rodríguez-Girondo and Uña-Álvarez (2016)[15] also introduced methods for checking the Markov assumption for the progressive illness-death model. The performance of their methods was studied through simulation studies. Among the methods for simulating data, their model 2 is the one that we aim to reproduce in our scenario 4, making some comparisons possible. As in their case, our simulations reveal the inability of the Cox model to identify the failure of the Markovianity with proportion rejections varying between 5%

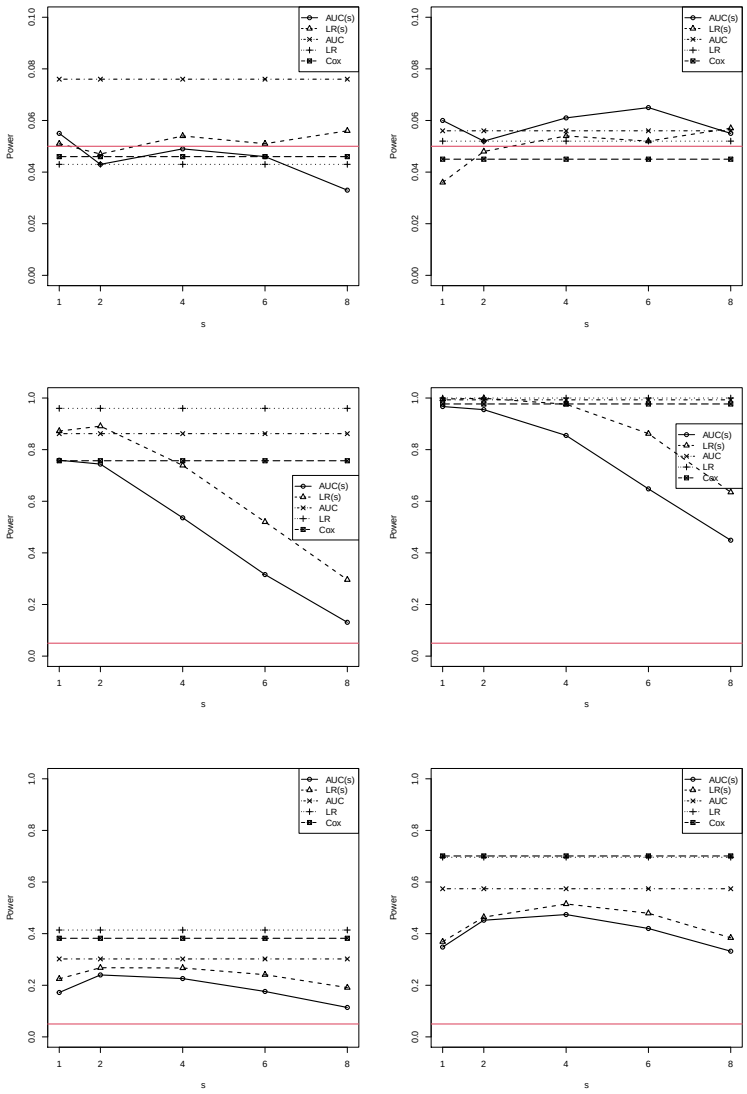

Fig. 2 Rejection probabilities for testing the null hypothesis of the Markov condition for the three tests for nominal level 5%. Markov, semi-Markov and non-Markov scenarios (upper, middle, and lower panels, respectively), for $n = 250$ and $n = 500$ (left and right panels, respectively). Results for the transition probability $\hat{p}_{23}(s,t)$. Censoring times uniformly distributed between 0 and 30.

and 10%. As in our case, the methods proposed in Rodríguez-Girondo and Alvarez (2016) revealed an increased power of the global tests as the sample size increases and with a decrease in the censoring percentage. Among the proposed tests, the wC_n method, based on the local Kendall's tau τ_i, appears to be the one with better accuracy to distinguish the non-markovianity of the process either for subjects who pass directly from State 1 to State 2 or for those that have passed through the intermediate state. Comparing the results of the AUC

Table 2 Rejection proportions for nominal level of 5% of the local tests for fixed values $s = 0.2$, $s = 0.6$, $s = 1$, $s = 1.2$, $s = 1.4$ and $s = 1.6$ (AUC(s) and LR(s)). Rejection proportions for the global tests (AUC, LR and Cox) are also included. Non-Markovian scenario, hazard with a quadratic predictor.

									Global	
Scenario	Trans. Prob.	n	Method	0.2	0.6	1	1.4	1.6	AUC/LR	Cox
Non-Markov	$\widehat{p_{12}}(s,t)$	250	AUC(s)	0.270	0.260	0.042	0.186	0.256	0.360	0.074
quadratic		500	AUC(s)	0.492	0.504	0.094	0.278	0.468	0.708	0.094
predictor	$\widehat{p_{23}}(s,t)$	250	LR(s)	0.348	0.283	0.053	0.169	0.264	0.439	0.074
		500	LR(s)	0.638	0.542	0.065	0.299	0.489	0.815	0.094
$C \sim U[0, 8.1]$	$\widehat{p_{23}}(s,t)$	250	AUC(s)	0.324	0.314	0.064	0.128	0.162	0.430	0.074
		500	AUC(s)	0.538	0.532	0.010	0.228	0.376	0.742	0.094
Non-Markov	$\widehat{p_{12}}(s,t)$	250	AUC(s)	0.250	0.276	0.072	0.092	0.186	0.410	0.092
quadratic		500	AUC(s)	0.422	0.455	0.112	0.122	0.256	0.638	0.107
predictor	$\widehat{p_{23}}(s,t)$	250	LR(s)	0.294	0.257	0.063	0.115	0.161	0.305	0.092
		500	LR(s)	0.535	0.449	0.059	0.172	0.287	0.647	0.107
$C \sim U[0, 4.6]$	$\widehat{p_{23}}(s,t)$	250	AUC(s)	0.238	0.294	0.080	0.062	0.098	0.420	0.092
		500	AUC(s)	0.416	0.430	0.114	0.094	0.168	0.642	0.107

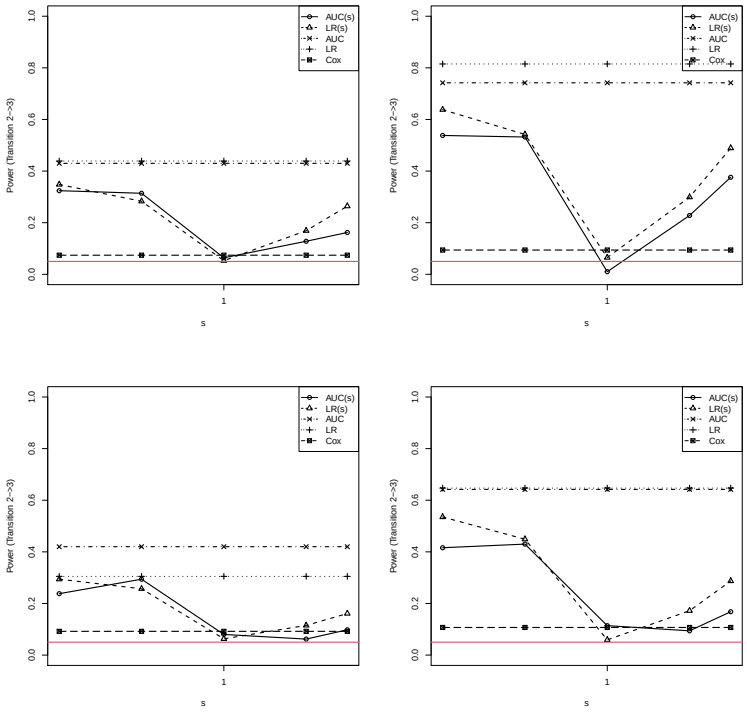

Fig. 3 Rejection probabilities for testing the null hypothesis of the non-Markov condition for the three tests for nominal level 5%. Non-Markovian scenario, hazard with a quadratic predictor. Results based on different censoring percentages ($C \sim U[0, 8.1]$ - upper, $C \sim U[0, 4.6]$ - bottom), for $n = 250$ and $n = 500$ (left and right panels, respectively). Results for the transition probability $\hat{p}_{23}(s,t)$.

global test, reported in our Table 2, to the proposed wC_n method, namely for individuals that experienced a transition through the intermediate state, we can observe higher rejection proportions for the AUC test for all samples sizes (n=250 and n=500) and censoring parameters (4.6 and 8.1). It is worth remember that the extension of the methods proposed in

[15] to general models is not straightforward, while our methods (based on the AUC) can be applied to general multi-state models as illustrated in our third real data example.

4 Real data analysis

In this section we illustrate the proposed methods using data from three clinical trial studies. We first use data from a colon cancer study from a large clinical trial on Duke's stage III patients (Moertel *et al.*, 1995) [22], the second one is from a clinical trial on breast cancer and the last one from a data set of liver cirrhosis patients subjected to a prednisome treatment (Andersen *et al.*, 1993) [1].

Surgical resection is the best treatment option for cancer patients and the most powerful tool for assessing prognosis following potentially curative surgery. In a large percentage of the patients with such cancers, 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 the disease and death (in some cases). Cancer patients who have experienced a recurrence are known to be at a substantially higher risk of mortality. Usually, this mortality is higher in cases of early recurrences. The effect of a recurrence in a survival model is traditionally studied using extensions of the Cox proportional hazards model (Cox, 1972 [23]; Genser and Wernecke, 2005 [24]). Multi-state models can also be successfully used to model such data (Pérez-Ocón *et al.*, 2001 [25]; Putter *et al.*, 2007 [26]; Meira-Machado et al., 2009 [3]; Meira-Machado, 2016 [27]; Meira-Machado e Sestelo, 2019 [4]). In both real data examples from cancer studies, data can be viewed as arising from a progressive illness-death model with states 'Alive and disease-free', 'Alive with Recurrence' and 'Dead'. Below, the Markov assumption is carefully analyzed comparing the proposed methods with the traditional approach.

4.1 Colon cancer study

In this study, 929 patients affected by colon cancer were followed from the date of a curative surgery for colorectal cancer until censoring or death from colon cancer. From this total, 468 developed a recurrence and among these 414 died; 38 patients died without recurrence. The rest of the patients (423) remained alive and disease-free up to the end of the follow-up.

Figure 4 reports estimated transition probabilities for fixed values of $s = 365, 730, 1095$ and 1460 days (1, 2, 3 and 4 years, respectively), along time, for the transition probabilities $\widehat{p}_{12}(s,t)$ (left hand side) and $\widehat{p}_{23}(s,t)$ (right hand side). As expected, these plots reveal that the landmark estimators (`LM` and `LMAJ`) have more variability than the Aalen-Johansen estimator (`AJ`). This is a obvious consequence of the subsampling approach which will be more evident for some specific values of s and higher values of t. Plots shown in the top of the figure (for s equal to one year) show departures between the two Markov-free estimators (`LM` and `LMAJ`) and the Aalen-Johansen estimator (`AJ`). Note that for the mortality transition from State 2 to State 3 the two (Markov-free) landmark estimators are equivalent. Deviations from the two approaches (Markovian and Markov-free), as those shown for s equal to one year, may be explained by the failure of the Markov assumption. On the other hand, the corresponding plots for the remaining values of s show that all methods behave quite similar.

Plots shown in the first row of Figure 5 compare the Aalen-Johansen estimator (`AJ`) and the landmark non-Markovian estimator (`LMAJ`) for $p_{12}(s = 365, t)$, $p_{13}(s = 365, t)$ and $p_{23}(s = 365, t)$. A small deviation can be seen in these plots with respect to the

Table 3 Probability values of the local test for several fixed values of s (measured in days). Rejection proportions for the global tests also included. Colon cancer data.

								Global	
Trans. Prob.	Method	90	180	365	730	1095	1460	AUC / LR	Cox
$\widehat{p}_{12}(s,t)$	AUC(s)	0.012	0.007	0.002	0.154	0.135	0.857	0.014	0.154
$\widehat{p}_{23}(s,t)$	LR(s)	0.006	0.026	0.036	0.685	0.981	0.509	0.018	0.154
	AUC(s)	0.003	0.004	0.003	0.155	0.118	0.714	0.013	0.154

straight line $y = x$. The plot on the second row presents the estimated transition probabilities $\widehat{p}_{23}(s = 365, t)$ from the landmark Aalen-Johansen estimator with 95% pointwise confidence limits (black lines) and Aalen-Johansen estimator (red line), revealing some discrepancies between the two approaches in the estimation of this transition probability. These plots provide a graphical test of the Markov assumption which reveal some evidence on the lack of Markovianity of the underlying process beyond one year after surgery.

For further illustration, in Figure 6 we display the discrepancy between the Aalen-Johansen estimator (Markovian) and the landmark non-Markovian estimator (`LMAJ`), for $p_{12}(s,t)$ and $p_{22}(s,t)$, for $s = 365$, $s = 730$, $s = 1095$ and $s = 1460$, measured through $D_{hj} = \widehat{p}_{hj}^{\,\mathtt{AJ}}(s,t) - \widehat{p}_{hj}^{\,\mathtt{LMAJ}}(s,t)$, $h = 1, 2$, $j = h + 1$. The 95% pointwise confidence limits were obtained using simple bootstrap. This plot reveals clear differences between the two methods in large intervals for $s = 365$. The differences are observed by the deviation of the plot with respect to the straight line $y = 0$, from which one gets some evidence on the lack of Markovianity of the underlying process beyond one year after surgery. On the other hand the plots depicted for other values of s do not reveal evidence against the Markov assumption. In summary, these plots show that there is some evidence, at least for $s = 365$, that the application of the Aalen-Johansen method is not recommended here, due to possible biases. They also reveal a possible failure of the Markov assumption. It is worth mention that deviations of the plots with respect to the straight line $y = 0$ in the right tail (higher values of t) should not be overvalued since they often occur due to the limited number of individuals at these times. Note that the findings observed in Figure 5 are not in agreement with the results obtained through the 'global' test for Markovianity based on the Cox model (using time to recurrence as a time-dependent covariate). This test reported a coefficient of negative sign for the recurrence time, according to an increased risk of death shortly after relapse (P-value = 0.154) revealing no evidence against the Markov model for the colon data.

Results reported in Table 3 are in agree with those obtained from the graphical inspection shown in Figure 6, revealing a failure of the Markov assumption only for non-null lower values of s. They show that, the test based on the difference of the area under the two curves lead to a probability value of 0.002 and 0.003, respectively for $\widehat{p}_{12}(s,t)$ and $\widehat{p}_{23}(s,t)$, for $s = 365$. Low probability values (less than 5%) were also obtained for $s = 90$ and $s = 180$ too. The global test we propose (based on the areas under the transition probabilities) are also in agreement with our findings, reporting a probability value lower than 0.014 against the Markov condition. The local tests based on the log-rank statistic also confirmed small probability values mainly for s up to 365. Either the AUC and the log-rank global tests confirm the failure of the Markovianity of the process.

Often multi-state models include covariates and it may be the omission of covariate effects that induces apparent non-Markovianity. The methods proposed in this paper can also deal with this problem since discrete covariates can be included in the estimation of the transition probabilities $p_{hj}(s,t)$ by splitting the sample for each level of the covariate and repeating the described procedures for each subsample. As shown in Table 4 treat-

Table 4 Probability values of the local test for $s = 365$ days by treatment for AUC local test. Rejection proportions for the test based on the Cox model also included. Colon cancer data.

Trans. Prob.	Treatment	Method	s=365	Cox
$\hat{p}_{12}(s,t)$	Obs	AUC(s)	0.0002	
	Lev		0.7192	
	Lev+5FU		0.1116	
$\hat{p}_{23}(s,t)$	Obs	AUC(s)	0.0008	0.062
	Lev		0.3013	0.401
	Lev+5FU		0.1562	0.712

ment (Obs(ervation), Lev(amisole), Lev(amisole)+5-FU) revealed a strong effect on the 2→3 transition intensities and a greater effect on 1→ 2. Results reported in Table 4 also show that the test for Markovianity based on the Cox model reported a p value of 0.062 (regression coefficient: -0.000528) for the Observation group.

4.2 Breast cancer data

In this section we use data from the second trial in which a total of 720 women with primary node positive breast cancer is recruited in the period between July 1984 and December 1989. The data which was also used by Sauerbrei and Royston (1999) [28] considers 686 patients who had complete data for the two event times (time to recurrence and time to death). In this study, patients were followed from the date of breast cancer diagnosis until censoring or dying from breast cancer. From the total of 686 women, 299 developed a recurrence and 171 died.

As for the analysis of the colon cancer data, we start to present on Figure 7 the estimated transition probabilities for fixed values of $s = 365$, 730, 1095 and 1460 days, along time, for the transition probabilities $\widehat{p}_{12}(s,t)$ (left hand side) and $\widehat{p}_{23}(s,t)$ (right hand side). In this case, differences between the estimated curves of the Aalen-Johansen (AJ) and the Landmark estimator (LMAJ) are not evident. The discrepancy of the two estimators with the 95% pointwise confidence limits is also displayed in Figure 8 for $D_{hj} = \widehat{p}_{hj}^{\,\mathtt{AJ}}(s,t) - \widehat{p}_{hj}^{\,\mathtt{LMAJ}}(s,t), h = 1,2, j = h+1$. In this case, there are no clear evidences of a deviance of the plot with respect to the straight line $y = 0$, at least in large intervals. In summary, these plots do not show evidence against the use of the Aalen-Johansen estimator and therefore, against the Markov assumption. These findings are in agree with the results obtained through the three 'global' tests for Markovianity in Table 5. The test based on the Cox model which reported a coefficient of negative sign for the recurrence time, according to an increased risk of death shortly after relapse (P-value = 0.121) revealing no evidence against the Markov model for the breast cancer data. Higher probability values were obtained from the global test based on the area under the transition probabilities and log-rank statistics. The two local tests confirm this fact too.

Table 5 Probability values of the local test for several fixed values of s (measured in days). Rejection proportions for the global tests also included. Breast cancer data.

							Global	
Trans. Prob.	Method	180	365	730	1095	1460	AUC / LR	Cox
$\hat{p}_{12}(s,t)$	AUC(s)	0.543	0.306	0.232	0.247	0.241	0.230	0.121
$\hat{p}_{23}(s,t)$	LR(s)	0.926	0.647	0.246	0.163	0.922	0.580	0.121
	AUC(s)	0.955	0.603	0.269	0.428	0.577	0.280	0.121

4.3 Liver cirrhosis data

In this section we consider a data set of liver cirrhosis patients who were included in a randomized clinical trial at several hospitals in Copenhagen between 1962 and 1974. The study aimed to evaluate whether a treatment based on prednisone prolongs survival for patients with cirrhosis [1]. Let State 1 correspond to 'normal prothrombin level', State 2 to 'low (or abnormal) prothrombin level', and the State 3 to 'dead'. The movement of the patients among these three states can be modeled through the reversible multi-state model shown in Figure 9. From the total of 488 patients with liver cirrhosis initially enrolled in the study, 292 ended up to died, from which 104 experienced a direct transition from State 1 to the absorbing state, and in 188 patients an abnormal prothrombin level was detected at any time. There were also 314 patients that had movements from abnormal prothrombin levels towards normal levels and 274 from the normal prothrombin level to the intermediate state. Most transition times are below 1460 days, with a maximum of 4892 days.

Following the same procedure of the previous real data set analysis, we started comparing the estimated curves of the LMAJ and the AJ estimators for the transitions probabilities $\widehat{p}_{12}(s,t)$, $\widehat{p}_{21}(s,t)$ and $\widehat{p}_{23}(s,t)$, for fixed values of s= 180, 365, 730 and 1095 days, with the purpose to identify a possible failure of the Markov assumption. These times were chosen to cover the first years of the study corresponding to the most cases with transitions. In fact, after 4 years for all transitions the number of individuals decrease with potential consequences for the estimates under the landmark approach as referred previously in case of small size samples. The plots with the estimated curves at those points are shown in Figure 10. Plots shown in the first column reveal some departures between LMAJ and AJ estimators of $p_{12}(s,t)$, but only for lower values of s. The deviation between the two estimators seem to be more evident when comparing the estimated curves of $p_{21}(s,t)$ (second column), while this is not so evident when comparing the estimated curves of the transition probability $p_{23}(s,t)$ (third column). As referred above, apparent deviation between the two estimated curves, at least at some lowers values of s, may due to the lack of Markov condition.

The discrepancy of the two estimators, computed using $D_{hj} = \widehat{p}_{hj}^{\,\mathtt{AJ}}(s,t) - \widehat{p}_{hj}^{\,\mathtt{LMAJ}}(s,t)$ with the 95% pointwise confidence limits is also displayed in Figure 11. Some of these plots reveal some evidence of a deviance of the plot with respect to the straight line $y = 0$, revealing a possible failure of the Markov condition. Some of these findings are in agreement with the results reported in Table 6, which shows the rejection proportions, for $\widehat{p}_{12}(s,t)$, $\widehat{p}_{21}(s,t)$ and $\widehat{p}_{23}(s,t)$ of the proposed tests for checking the Markov assumption. Results were obtained by the empirical rejection proportions from 250 trials at the significant level of 0.05. Interestingly, the proposed local test was able to detect a failure of the Markov condition for $s = 365$ for the mortality transition of patients with abnormal prothrombin level. For the remaining time points of s, the test obtained lower rejection probabilities which are in agreement with the results obtained in all global tests. For the transition from State 1 to State 2, the proposed local test only reveal the failure of the Markov condition for $s = 180$. For the transition 2 to 1, besides $s = 180$, the local test also revealed a failure of the Markov condition for $s = 365$. These evidences (of failure of the Markov condition) for these two transitions are confirmed by the results of the proposed global test based on the AUC and the test based on the Cox model.

Table 6 Probability values of the local test for several fixed values of s (measured in days). Rejection proportions for the global tests also included. Liver cirrhosis data.

							Global	
Trans. Prob.	Method	180	365	730	1095	1460	AUC / LR	Cox
$\hat{p}_{12}(s,t)$	AUC(s)	0.002	0.158	0.134	0.639	0.793	$<$0.001	0.002
$\hat{p}_{21}(s,t)$	AUC(s)	$<$0.001	$<$0.001	0.156	0.253	0.237	0.001	$<$0.001
$\hat{p}_{23}(s,t)$	LR(s)	0.699	0.336	0.594	0.641	0.034	0.298	0.999
$\hat{p}_{23}(s,t)$	AUC(s)	0.317	0.030	0.677	0.367	0.195	0.258	0.999

5 Discussion

The Markov assumption is commonly used to analyze multi-state survival data. Therefore, goodness-of-fit tests for the Markov assumption are crucial in these models. Traditionally, this assumption is tested including covariates depending on the history on the modeling process. The comparison between estimated transition probabilities is the basis to introduce two formal local tests for the Markov assumption. The new methods are based on measuring the discrepancy of the Aalen-Johansen estimator which gives consistent estimators in Markov processes, and recent approaches that do not rely on this assumption. A log-rank test is used on specific transitions to check if the Markov assumption holds. A second method is proposed in this paper in which the test statistic is based on the difference of the areas under the two curves. We note that alternative test statistics could also have been considered such as those based on the absolute differences or squared differences between the Aalen-Johansen and the landmark estimators that would lead to a kolmogorov smirnov or a Cramer-von Mises-type test statistic, respectively.

Simulation results reveals that the two methods perform similarly revealing high power to detect a failure of the Markov condition. The simulation results and the results obtained through real medical data analysis suggest that the second approach may be a good alternative to the existing methods. The use of the graphical local tests based on the discrepancy between estimated curves of the transition probabilities, proposed here, are recommended to confirm the conclusions obtained from the application of this formal local test. In general, the two curves may cross at mid time points when the process is indeed Markovian (and the two curves are similar). If the process is not Markovian, then it is expected that the two curves only cross at earlier time points or at higher time points (at the right tail). Nevertheless, it is wise to start the analysis with a graphical test in particular to identify possible situations in which the process is indeed not Markovian and the two curves cross at mid time points. In such cases the usage of a different test statistics (e.g. based on a squared difference) should be also analyzed in future research investigation.

The use of local tests is recommended whenever the interest is focused on the estimation of the transition probabilities and, in particular, to decide which estimator is the most appropriate to use: the Aalen-Johansen estimator or a robust estimator. The use of the proposed local test is advised for each transition probability $p_{hj}(s,t)$ ($h > 1$), and the use of the robust Markov-free estimator when faced of evidences against Markovianity. This procedure may be followed for a general multi-state model.

A global test, such as the test proposed here, might be preferable in regression purposes. To this end, a common simplifying strategy is to decouple the whole process into various survival models, by fitting separate intensities to all permitted transitions using semi-parametric Cox proportional hazard regression models, while making appropriate adjustments to the risk set. The most common models are characterized through one of the two model assumptions that can be made about the dependence of the transition intensities and time. The transition intensities may be modeled using separated Cox models assuming the process to

be Markovian (also known as the clock forward modeling approach). When the test rejects the Markov assumption, a possible alternative is to use a semi-Markov Cox model in which the future of the process does not depend on the current time but rather on the duration in the current state. Both models can be easily implemented using standard software such as the R packages `survidm` or `mstate`. To decide the appropriate modeling approach, the global test should be used to all transitions depending on history.

The global test proposed is obtained through the combination of the results from local tests over different times. Simulation results show that the proposed global test may be much more powerful than the standard parametric method based on the proportional hazard specification which relies on a prior model specification that may fail in practice. The proposed methods can be used in general multi-state models.

Discrete covariates can be included in the proposed methods by splitting the sample for each level of the covariate and repeating the described procedures for each subsample. To account for the effect of continuous covariates, one can consider estimators of the transition probabilities conditional on covariates. One standard method is to consider estimators based on a Cox's model fitted marginally to each type of transitions, with the corresponding baseline hazard function estimated by the Breslow's method.

We implemented all the proposed methods in R. The code in the form of an R package is available from the authors upon request.

Acknowledgements This research was financed by Portuguese Funds through FCT - "Fundação para a Ciência e a Tecnologia", within the research grants PTDC/MAT-STA/28248/2017 and PD/BD/142887/2018.

Conflict of Interest

The authors have declared no conflict of interest. (or please state any conflicts of interest)

References

1. Andersen, P.K. and Borgan, O. and Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.
2. Hougaard, P. (2000). *Analysis of multivariate survival data*. Springer-Verlag, New York.
3. Meira-Machado, L. and de Uña-Álvarez, J. and Cadarso-Suárez, C. and Andersen, P.K. (2009). Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research* **18**, 195–222.
4. Meira-Machado, L. and Sestelo, M. (2019). Estimation in the progressive illness-death model: A nonexhaustive review. *Biometrical Journal* **61**, 2, 245–263.
5. Andersen, P.K. and Esbjerg, S. and Sorensen, T.I.A. (2000). Multistate models for bleeding episodes and mortality in liver cirrhosis. *Statistics in Medicine* **19**, 587–599.
6. Andersen, P.K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research* **11**, 91–115.
7. Aalen, O. and Johansen, S. (1978). An Empirical transition matrix for non homogeneous Markov and chains based on censored observations. *Scandinavian Journal of Statistics* **5**, 141–150.
8. Meira-Machado, L. and de Uña-Álvarez, J. and Cadarso-Suárez, C. (2006). Nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis* **12**, 325–344.
9. Allignol, A., Beyersmann, J., Gerds, T. and Latouche, A. (2014). A competing risks approach for nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis* **20**, 495–513.
10. Titman, A.C. (2015). Transition probability estimates for non-Markov multi-state models. *Biometrics* **71**, 1034–1041.
11. de Uña-Álvarez, J. and Meira-Machado, L. (2015). Nonparametric Estimation of Transition Probabilities in the Non-Markov Illness-Death Model: A Comparative Study. *Biometrics* **71**, 364–375.
12. Putter, H. and Spitoni, C. (2018). Non-parametric estimation of transition probabilities in non-Markov multi-state models: The landmark Aalen-Johansen estimator. *Statistical Methods in Medical Research* **27**, 7, 2081–2092.

13. Kay, R. (1986). A Markov model for analyzing cancer markers and disease states in survival studies. *Biometrics* **42**, 457–481.
14. Rodriguez-Girondo, M. and de Uña-Álvarez, J. (2012). A nonparametric test for Markovianity in the illness-death model. *Statistics in Medicine* **31**, 30, 4416–4427.
15. Rodriguez-Girondo, M. and de Uña-Álvarez, J. (2016). Methods for testing the Markov condition in the illness-death model: a comparative study. *Statistics in Medicine* **35**, 20, 3549–3562.
16. Titman, A.C. and Putter, H. (2020). General tests of the Markov property in multi-state models. *Biostatistics*. DOI: 10.1093/biostatistics/kxaa030.
17. Chiou, S.H., Qian, J., Mormino, E., Betensky, R.A. (2018). Permutation tests for general dependent truncation. *Computational Statistics & Data Analysis*. **128**, 308–324.
18. Beyersmann, J. and Schumacher, M. and Allignol, A. (2012). *Competing Risks and Multistate Models with R*. Springer, New York.
19. Borgan, O. (2005). Encyclopedia of biostatistics: Aalen-Johansen estimator. John Wiley & Sons.
20. Datta, S. and Satten, G.A. (2001). Validity of the Aalen-Johansen estimators of stage occupation probabilities and Nelson-Aalen estimators of integrated transition hazards for non-Markov models. *Statistics & Probability Letters* **55**, 403–411.
21. Van Houwelingen, H. C. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics* **34**, 1, 70–85.
22. Moertel, C.G. and Fleming, T.R. and Macdonald, J.S. and Haller, D.G. and Laurie, J.A. and Tangen, C.M. and Ungerleider, J.S. and Emerson, W.A. and Tormey, D.C. and Glick, J.H. and Veeder, M.H. and Mailliard, J.A. (1995). Fluorouracil Plus Levamisole as Effective Adjuvant Therapy after Resection of Stage III. Colon Carcinoma: A Final Report. *The Annals of Internal Medicine* **122**, 5, 321–326.
23. Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–200.
24. Genser, B. and Wernecke, K.D. (2005). Joint Modelling of Repeated Transitions in Follow-up Data – A Case Study on Breast Cancer Data. *Biometrical Journal* **47**, 3, 388–401.
25. Pérez-Ocón, R. and Ruiz-Castro, J.E. and Gámiz-Pérez, ML (2001). Non-homogeneous Markov models in the analysis of survival after breast cancer. *Journal of the Royal Statistical Society, Series C* **50**,1, 111–124.
26. Putter, H. and Fiocco, M. and Geskus, R.B. (2007). Tutorial in biostatistics: competing risks and multistate models. *Statistics in Medicine* **26**, 2389–2430.
27. Meira-Machado, L.(2016). Smoothed landmark estimators of the transition probabilities. *SORT-Statistics and Operations Research Transactions* **40**, 375–398.
28. Sauerbrei, W. and Royston, P. (1999). Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society, Series A* **161**, 1, 71–94.

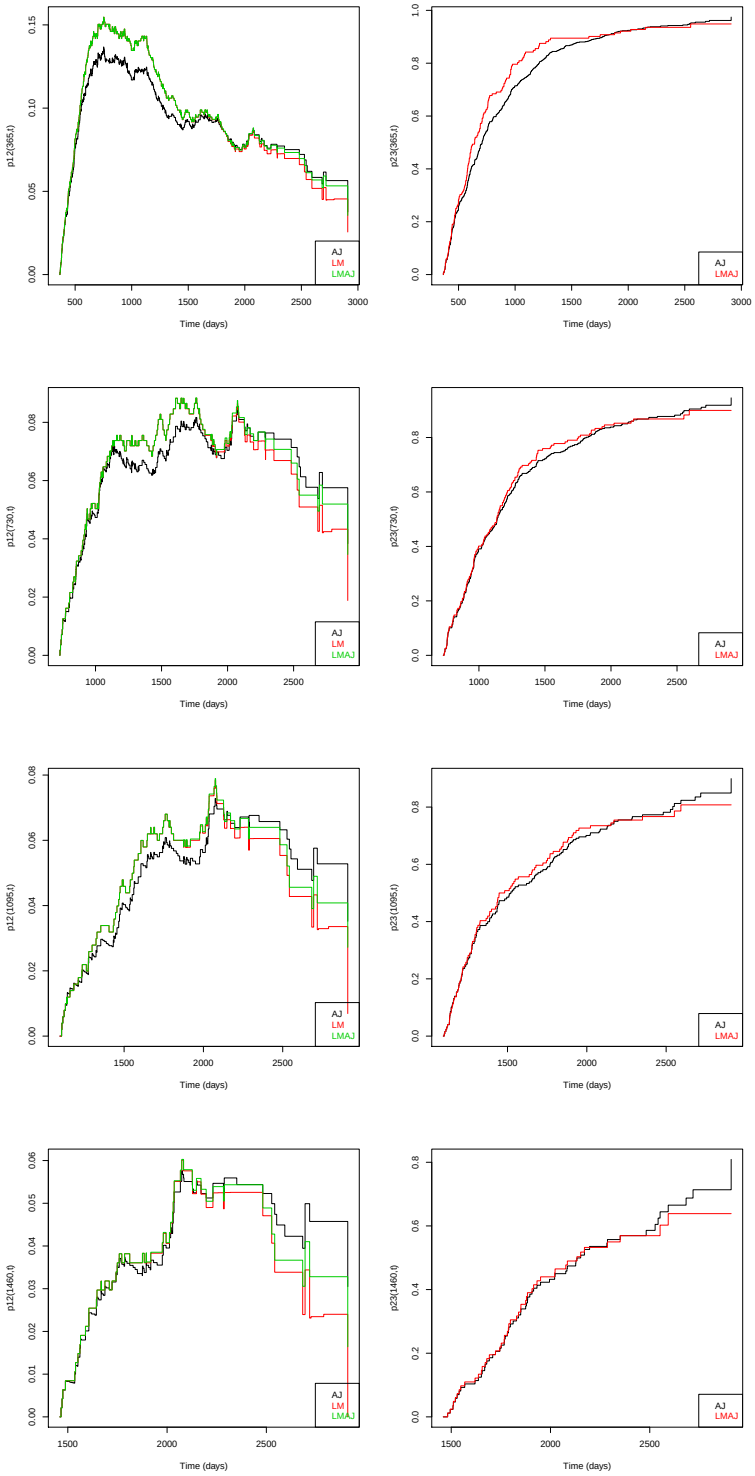

Fig. 4 Estimates of the transition probabilities for the Aalen-Johansen (AJ) and Markov-free estimators (landmark and landmark Aalen-Johansen), for s equal to 1, 2, 3 and 4 years since entry in study. Colon cancer data. Transition probabilities of $\widehat{p_{12}}(s,t)$ (Left column) and $\widehat{p_{23}}(s,t)$ (Right column).

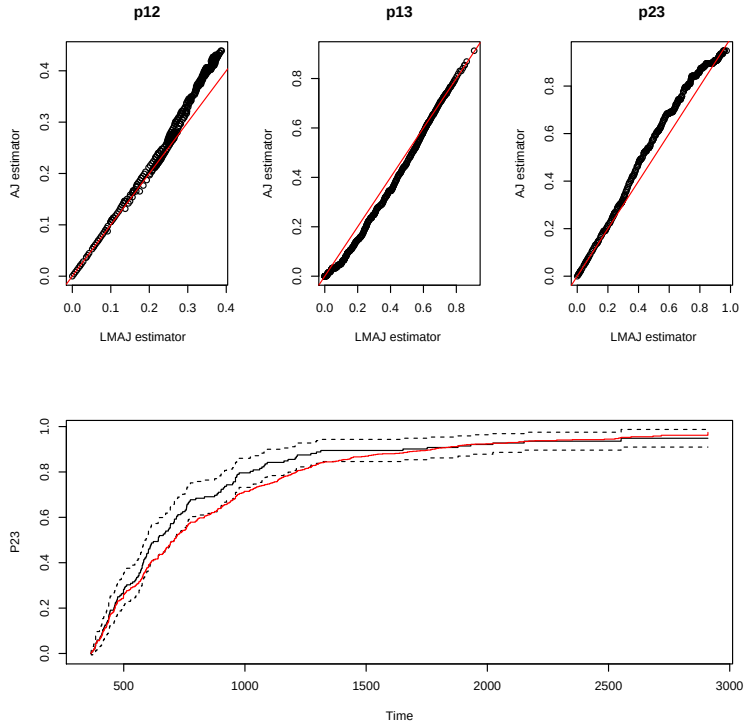

Fig. 5 Graphical test for the Markov condition, $s = 365$ (First row). Transition probabilities of $\widehat{p}_{23}(s = 365, t)$ from the landmark Aalen-Johansen estimator with 95% pointwise confidence limits (black lines) and Aalen-Johansen estimator (red line) (Second row). Colon cancer study.

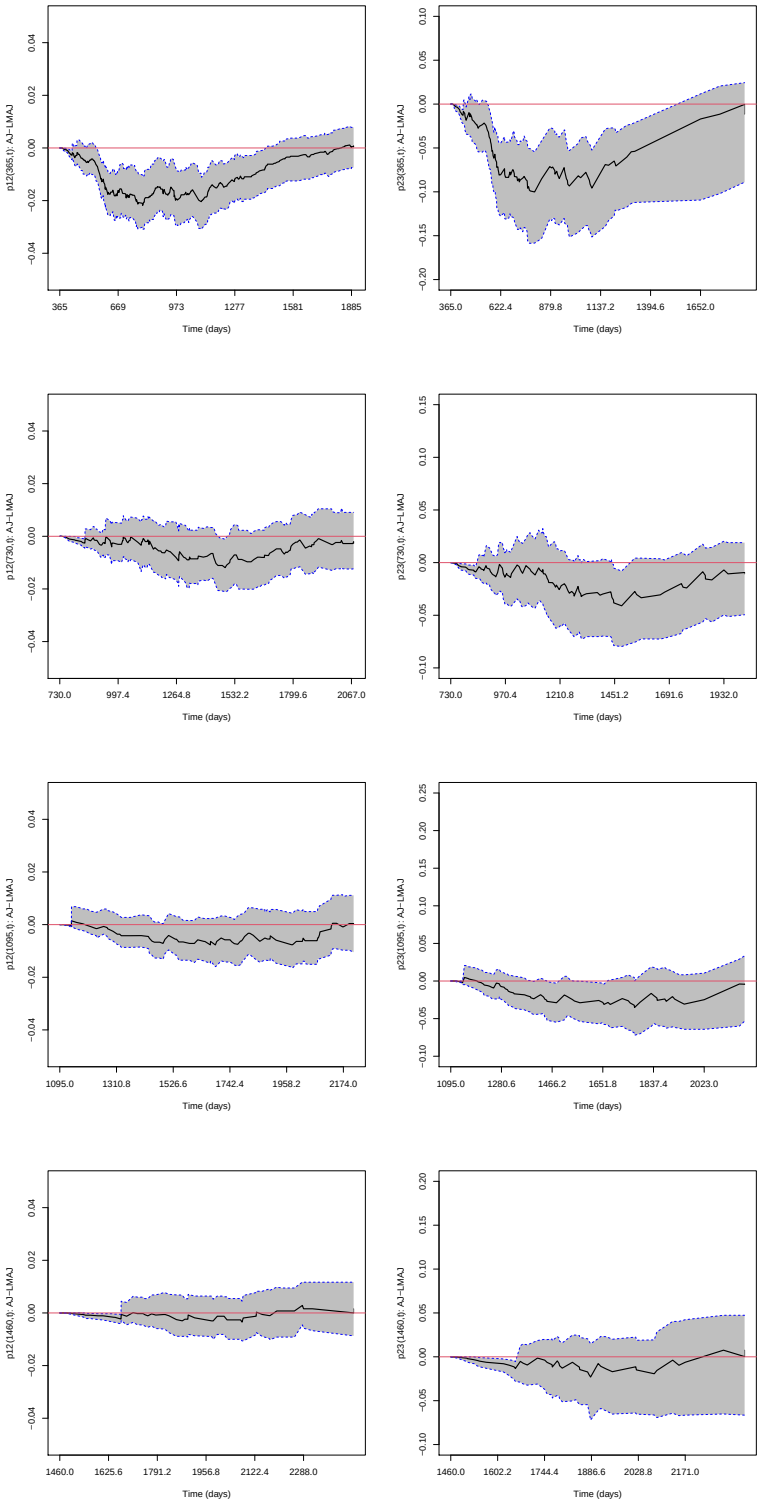

Fig. 6 Local graphical test for the Markov condition, for s equal to 1, 2, 3 and 4 years since entry in study. Test based on the discrepancy between the Aalen-Johansen estimator (Markovian) and the Markov-free estimator (`LM`). Colon cancer data.

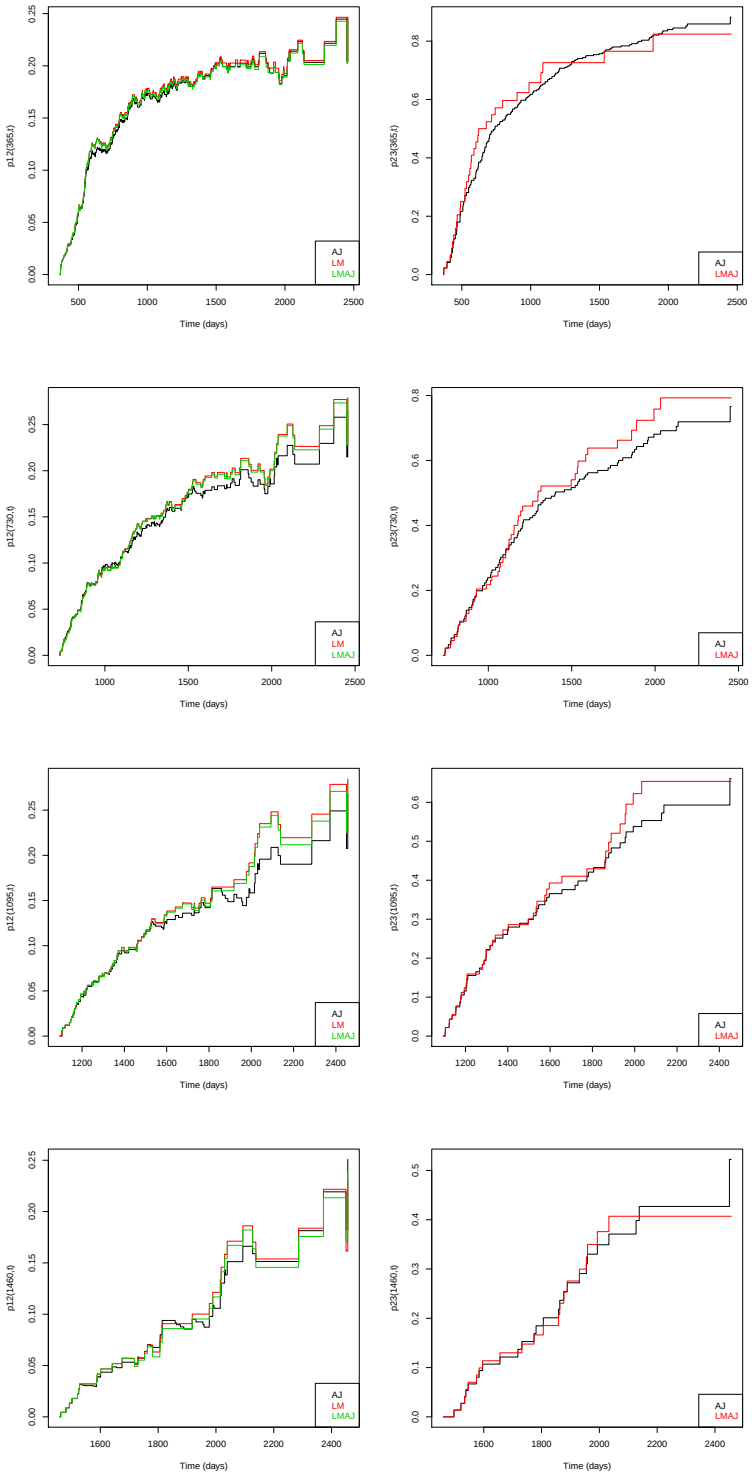

Fig. 7 Estimates of the transition probabilities for the Aalen-Johansen (AJ) and Markov-free estimators (landmark and landmark Aalen-Johansen), for s equal to 1, 2, 3 and 4 years since entry in study. Breast cancer data.

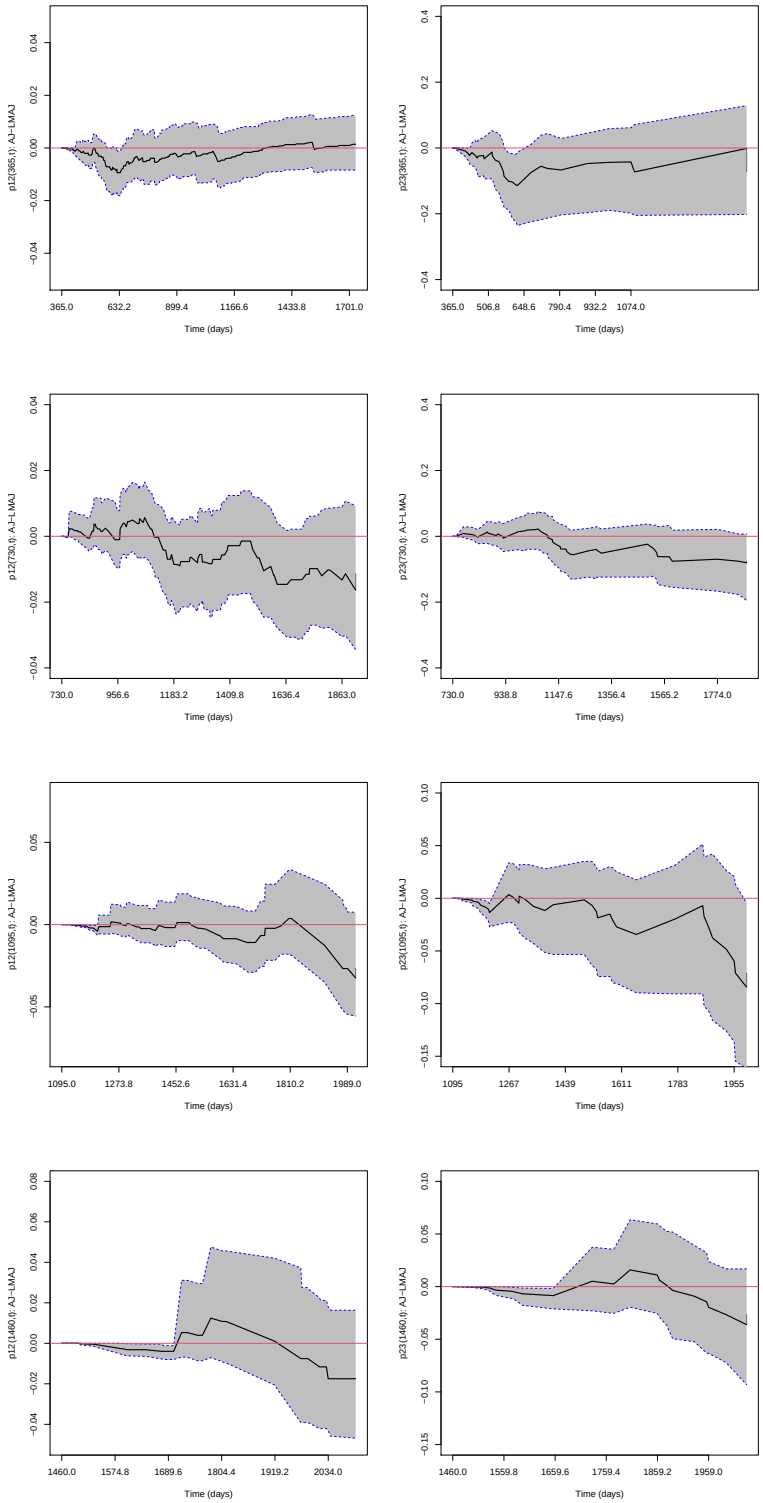

Fig. 8 Local graphical test for the Markov condition, for s equal to 1, 2, 3 and 4 years since entry in study. Test based on the discrepancy between the Aalen-Johansen estimator (Markovian) and the Markov-free estimator (`LM`). Breast cancer data.

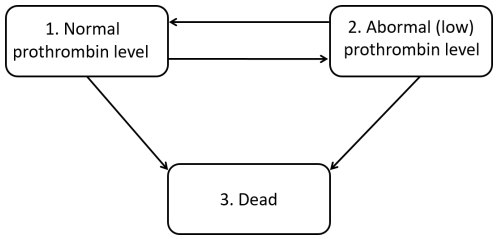

Fig. 9 The reversible illness-death model for patients with liver cirrhosis.

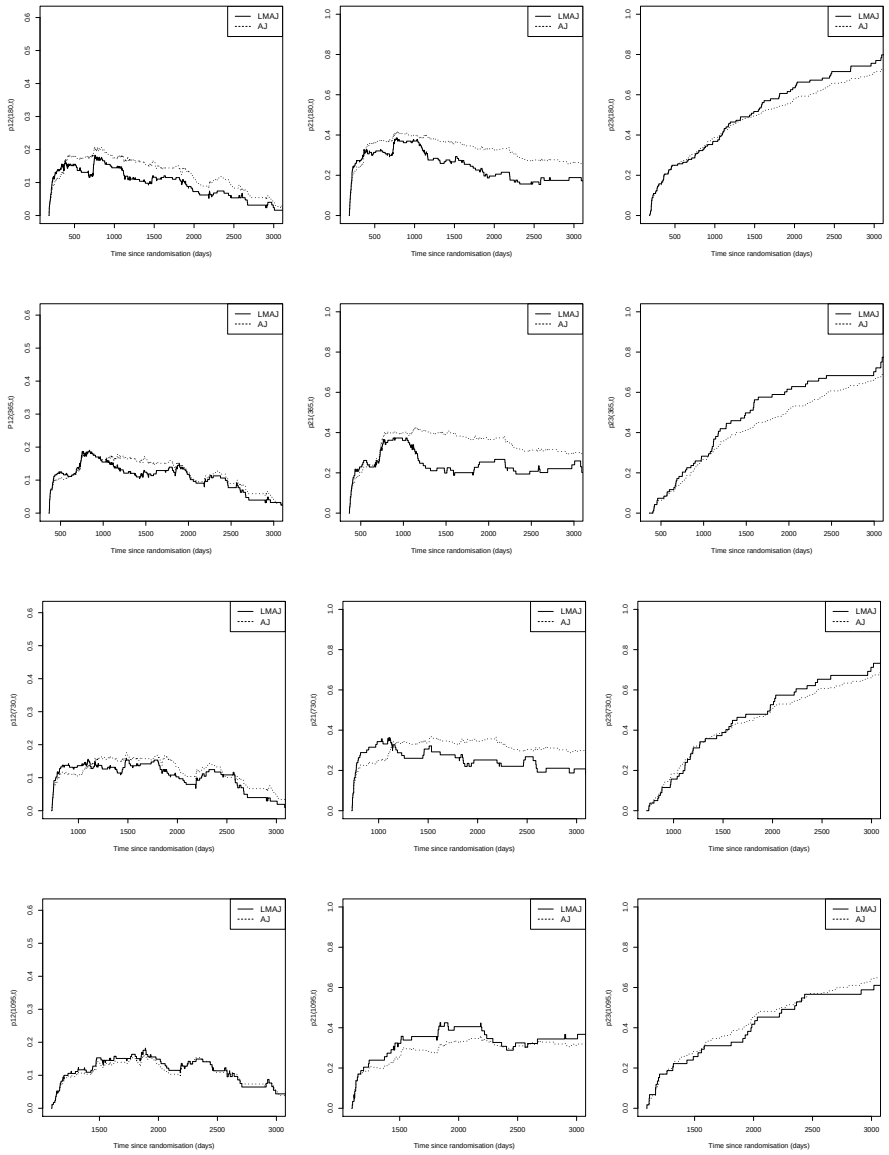

Fig. 10 Estimates of the transition probabilities for the Aalen-Johansen (AJ) and Markov-free estimators (landmark Aalen-Johansen), for some s equal to 180, 365, 730 and 1095 days since entry in study. Liver cirrhosis data.

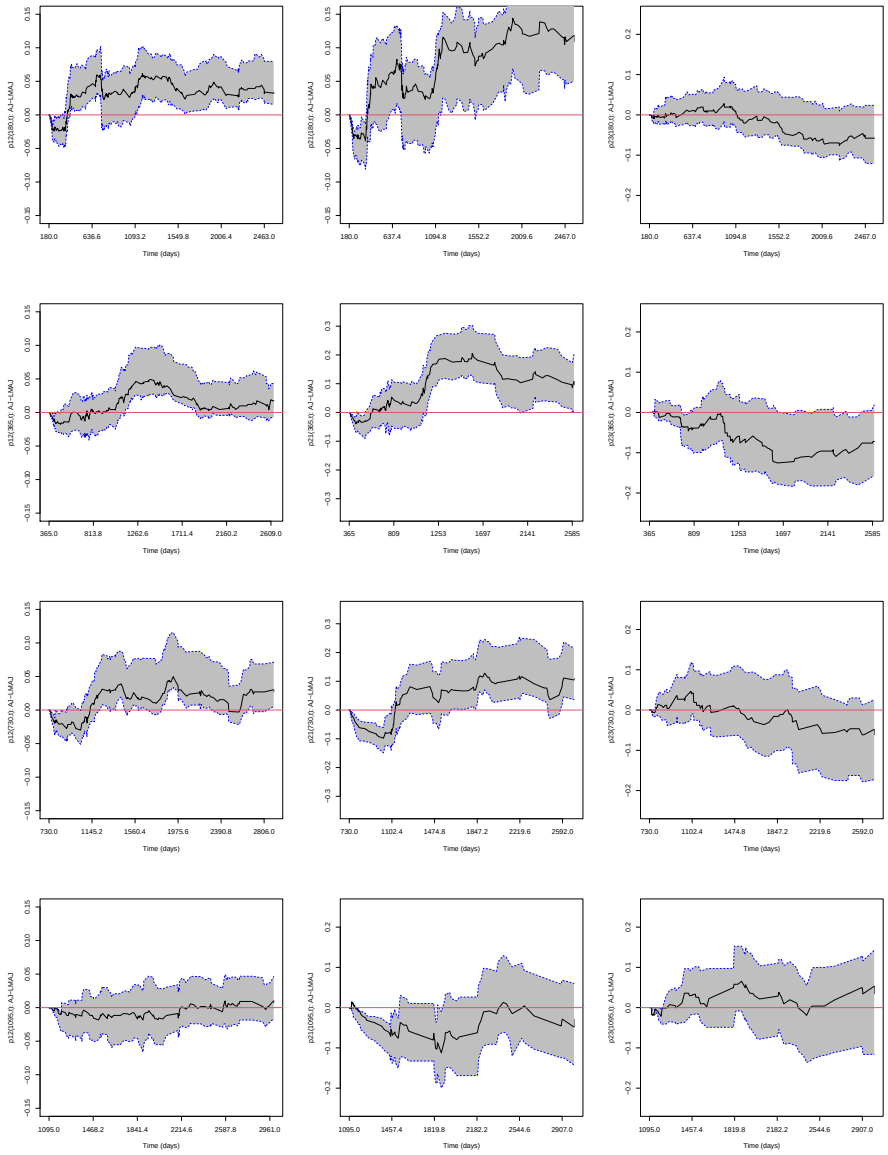

Fig. 11 Local graphical test for the Markov condition, for s equal to 180, 365, 730 and 1095 days since entry in study since entry in study. Test based on the discrepancy between the Aalen-Johansen estimator (Markovian) and the Markov-free estimator (`LMAJ`). Liver cirrhosis data.