General tests of the Markov property in multi-state models

ANDREW C. TITMAN[1] and HEIN PUTTER[2]

[1]Department of Mathematics and Statistics, Lancaster University, Lancaster, UK

[2]Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, , The Netherlands

July 9, 2020

Abstract

Multi-state models for event history analysis most commonly assume the process is Markov. This article considers tests of the Markov assumption that are applicable to general multi-state models. Two approaches using existing methodology are considered; a simple method based on including time of entry into each state as a covariate in Cox models for the transition intensities and a method involving detecting a shared frailty through a stratified Commenges-Andersen test. In addition, using the principle that under a Markov process the future rate of transitions of the process at times $t > s$ should not be influenced by the state occupied at time s, a new class of general tests is developed by considering summaries from families of log-rank statistics where patients are grouped by the state occupied at varying initial time s. An extended form of the test applicable to models that are Markov conditional on observed covariates is also derived. The null distribution of the proposed test statistics are approximated by using wild bootstrap sampling. The approaches are compared in simulation and applied to a dataset on sleeping behaviour.

The most powerful test depends on the particular departure from a Markov process, although the Cox-based method maintained good power in a wide range of scenarios. The proposed class of log-rank statistic based tests are most useful in situations where the non-Markov behaviour does not persist, or is not uniform in nature across patient time.

1 Introduction

Multi-state models provide a flexible framework for modelling event history data (Andersen and Keiding, 2002; Cook and Lawless, 2018). A multi-state model involves a stochastic process that can occupy one

of a number of possible states, which may for instance represent different stages of a chronic disease. The process is defined by the transition intensities, which govern the instantaneous risk of moving between states. In their most general form the transition intensities can be any function of the past history of the process. However, most commonly the process is assumed to be Markov or Markov conditional on observed covariates, meaning that the transition intensities are a function only of the current state occupied, time and possibly covariates, but no other part of the history. The advantage of assuming a Markov process is that, through the Kolmogorov forward equations, the transition probabilities can be linked to the transition intensities by a product limit relation. Non-parametric estimates of the transition probabilities can be obtained by applying that relation to non-parametric estimates of the transition intensities, leading to the Aalen-Johansen (AJ) estimator (Aalen and Johansen, 1978).

Recently there has been substantial interest in the development of non-parametric estimators of the transition probabilities in multi-state models that remain valid without requiring a Markovian assumption. Meira-Machado *and others* (2006); Allignol *and others* (2014); de Uña-Álvarez and Meira-Machado (2015) focussed on estimators for the three state progressive illness-death model. Titman (2015) proposed a general estimator based on constructing survival or competing risks processes. Putter and Spitoni (2018) proposed a simpler estimator, the landmark Aalen-Johansen estimator (LMAJ), which also has slightly better efficiency.

In practice, it is not generally possible to know in advance whether the Markov assumption is appropriate for a particular dataset. While the LMAJ may be applied, it will be less efficient and hence have a greater mean squared error than the standard AJ estimator in cases where the Markov assumption is valid or close to valid. Therefore, it is of practical importance to be able to test the Markov assumption.

For the progressive three-state illness-death model, Rodríguez-Girondo and de Uña-Álvarez (2012) have developed local and global tests of Markovianity based upon the observed Kendall's τ correlation between the time to illness and time to death among patients in the illness state at a given time. Recently Chiou *and others* (2018) developed tests for dependent truncation, allowing for the possibility of non-monotonic dependence that are also applicable to the progressive illness-death model. However, for more general multi-state models to our knowledge existing testing procedures are limited to *ad hoc* approaches, for instance including a specific aspect of the history, such as the entry time of a state of interest, in a model for the transition intensities (Kay, 1986). The stratified version of the score test of Commenges and Andersen (1995) can also be used to detect non-Markov behaviour induced by a shared frailty term affecting all, or a subset of the transition intensities.

The purpose of this paper is to develop methods for testing the Markov assumption that are applicable

to general multi-state models under right-censoring, and have adequate power under a range of biologically plausible violations of the Markov assumptions. The remainder of the paper is organized as follows: Section 2 provides general notation for the multi-state processes to be considered and outlines two existing approaches for general tests. Section 3 proposes a general test based on families of log-rank statistics. Section 4 presents a simulation study of the performance of the tests in a range of scenarios. Section 5 applies the tests to a data set on sleeping behaviour. The paper concludes with a discussion.

2 Approaches to general testing

Let $\{X_i(t), t \geq 0\}$ denote the multi-state process for subject $i = 1, \ldots, n$, where $X_i(t) \in \{1, \ldots, R\}$. Also let $Y_i(t)$ be the at risk indicator for the process. We assume all patients are observed from time 0. The transition intensities between states are defined as

$$\alpha_{lm}(t; \mathcal{H}_t) = \lim_{\delta t \downarrow 0} \frac{P(X(t + \delta t) = m | X(t-) = l, \mathcal{H}_t)}{\delta t},$$

for $l \neq m \in \{1, \ldots, R\}$, where $\mathcal{H}_t$ represents the history of $X(t)$ up to but not including time t. Note that direct transitions between some pairs of states may not be possible, in which case $\alpha_{lm}(t; \mathcal{H}_t)$ is zero for all times and past histories. Under a Markov assumption, the transition intensities simplify to $\alpha_{lm}(t; \mathcal{H}_t) = \alpha_{lm}(t)$. Covariates can be accommodated through proportional intensities assumptions such that

$$\alpha_{lm}(t; \mathbf{z}_i) = \alpha_{lm0}(t) \exp(\boldsymbol{\beta}_{lm} \mathbf{z}_i), \tag{2.1}$$

where $\mathbf{z}_i$ is a vector of explanatory variables for subject i and $\alpha_{lm0}(t)$ is a baseline intensity function. The case in which the baseline intensity functions are taken to be non-parametric functions of time is often referred to as a Cox-Markov model (Meira-Machado *and others*, 2009).

In the remainder of the section, two approaches to testing of the Markov assumption for general Markov or Cox-Markov models, using existing methodology are outlined.

2.1 Cox modelling approach

A simple to implement method of testing the Markov assumption is to include the most recent time of entry into state l, denoted t_l, as a covariate within Cox proportional hazard models for the transition intensities out of state l, such that $\alpha_{lm}(t; \mathcal{H}_t) = \alpha_{lm}(t; t_l) = \alpha_{lm0}(t) \exp(\theta_{lm} t_l)$, for a model that otherwise has no covariates and $\alpha_{lm}(t; \mathcal{H}_t, \mathbf{z}_i) = \alpha_{lm}(t; t_l, \mathbf{z}_i) = \alpha_{lm0}(t) \exp(\theta_{lm} t_l + \boldsymbol{\beta}_{lm} \mathbf{z}_i)$ for a model which is otherwise Cox-Markov (Kay, 1986; Meira-Machado *and others*, 2006).

Since $\alpha_{lm0}(t)\exp(\theta_{lm}t_l)$ can be rewritten as $\alpha_{lm0}(t)\exp(\theta_{lm}t)\cdot\exp(-\theta_{lm}(t-t_l))$, if the baseline hazard is unspecified and $\theta_{lm}\neq 0$, a non-homogeneous semi-Markov model arises where the transition intensities depend multiplicatively on both patient time t and time in the current state $t-t_l$. It is therefore reasonable to expect that the approach will be most effective if the violation of Markovianity is of this specific form. The Markov assumption is tested through a likelihood ratio test of the null hypothesis $\theta_{lm}=0$, to provide a test specific to a given transition, or $\boldsymbol{\theta}=\mathbf{0}$ for a global test across all transition intensities.

Throughout this paper we only consider using time of entry into the current state when using Cox modelling to test the Markov assumption. In principle, other summaries of the past history, such as the first time of entry into state l, or the number of previous visits to state l, can be used in addition to, or in place of t_l, to allow direct testing of other types of departure from a Markov process. However, choosing the appropriate summaries may be difficult without specific substantive knowledge of the process to be modelled.

2.2 Stratified Commenges-Andersen test

A non-Markov process can be induced by assuming there exist unobserved frailty terms, interpretable as the effects of unmeasured or unmeasurable explanatory variables, which are shared across the transitions of the multi-state model. Commenges and Andersen (1995) devised a score test of homogeneity for survival data. While this test was originally designed to detect heterogeneity in clustered survival data, the stratified version of the test, where in this case the different transition intensities are treated as separate strata, can be used as a test of homogeneity and hence the Markov assumption in Markov or Cox-Markov multi-state models. See Putter and van Houwelingen (2015) for more discussion on the role of frailties in "explaining" non-Markovianity in multi-state models.

In constructing the score test, the transition intensities for subject i are assumed to be of the form $\alpha_{lmi}(t;\mathbf{z}_i)=\alpha_{lm0}(t)\exp(\sigma\epsilon_i+\boldsymbol{\beta}_{lm}\mathbf{z}_i)$, where the ϵ_i are i.i.d. random variables with an unspecified distribution with mean 0 and variance 1. Hence a shared multiplicative frailty model with factor $\exp(\sigma\epsilon_i)$ is assumed. The test is then a score test of the null hypothesis $\sigma=0$, corresponding to zero variance, versus an alternative $\sigma>0$. Full details of the form of the test in the context of a multi-state model are given in Section S1 of the Supplementary Material.

The stratified version of the test is implemented within the **R** package frailtyEM (Balan and Putter, 2019). Note that this approach has the advantages of not requiring a frailty model to be fitted and not requiring a specific parametric form for the frailty distribution.

The assumption of the same frailty effect on every transition intensity is often not realistic. For instance,

we would not generally expect the same frailty effect to apply to forward direction transitions compared to backward transitions. The problem can be alleviated to some extent if a subset of transitions, $\mathcal{T}$, can be identified for which a similar frailty effect would be expected. The test statistic would then be based only on data from transitions in $\mathcal{T}$, where the formulae above remain unchanged except summations are with respect to l, m such that $(l, m) \in \mathcal{T}$.

While not pursued here, in principle the set of transition intensities could be organized into several disjoint subsets (e.g. forward transitions, backward transitions etc.). Under the null hypothesis of a Markov process, the statistics are asymptotically independent and standard normal and hence a χ^2 statistic can be devised to jointly test each subset of transitions.

3 Log-rank based statistics

3.1 Local test

The construction of the LMAJ motivates the development of alternative tests of the Markov assumption. Suppose initially that interest lies in assessing the validity of transition probability estimates from a specific start time s and state j. Two groups of subjects can be identified as $\mathcal{S} = \{i : X_i(s) = j, Y_i(s) = 1\}$ and $\mathcal{S}^c = \{i : X_i(s) \neq j, Y_i(s) = 1\}$. Under the LMAJ estimator, only the subjects in $\mathcal{S}$ would contribute to the estimate, whereas the AJ estimator would use both sets of subjects. Moreover, under a Markov process, the transition intensities of the process for $t > s$ will be the same in both groups. This property motivates a local test of Markovianity constructed by considering various log-rank statistics grouping by membership in $\mathcal{S}$.

The transition probabilities $\mathbf{P}_j(s,t) = (P_{j1}(s,t), \ldots, P_{jR}(s,t))'$ for $t > s$ are functions of the transition intensities $\alpha_{lm}(t)$ for $l \in \mathcal{R}_j$, where $\mathcal{R}_j$ represents the set of states that are reachable from state j. For each such transition intensity, interest lies in testing $H_0 : \alpha_{lm}(t|X(s) = j) = \alpha_{lm}(t|X(s) \neq j)$, for $t \geq s$ versus a general alternative. These hypotheses can be tested by defining $\delta_i^{(j)}(s) = I(X_i(s) = j)$ to be a group indicator with respect to the state of interest j. A log-rank statistic for each transition $l \to m$ is then of the form

$$U_s^{(j)}(l,m) = \sum_{i=1}^{n} \int_s^{\tau} \left\{ \delta_i^{(j)}(s) - \frac{\sum_k \delta_k^{(j)}(s) Y_{kl}(t)}{\sum_k Y_{kl}(t)} \right\} dN_i^{(lm)}(t), \tag{3.1}$$

where $Y_{il}(t) = I(X_i(t-) = l)Y_i(t)$ is the at risk indicator of transition $l \to m$ for subject i and τ is the maximum time of follow-up. Note that such a statistic will be uniformly 0 unless state l is also reachable if $X_i(s) \neq j$. Hence the test is defined for transitions from state $l \in \mathcal{R}_j \bigcap \mathcal{R}_{j^c}$ where $\mathcal{R}_{j^c} = \bigcup_{j' \neq j} \mathcal{R}_{j'}$, i.e.

transitions from states l that can be reached from state j and at least one other state $j' \neq j$.

The standardized statistics $\bar{U}_s^{(j)}(l,m) = U_s^{(j)}(l,m)/\sqrt{\widehat{\text{Var}}(U_s^{(j)})}(l,m)$, where $\widehat{\text{Var}}(U_s^{(j)})$ will be given in (3.3), can be compared to a $N(0,1)$ distribution to assess each individual hypothesis. Moreover, a local test of the Markov property can thus be constructed by combining the log-rank statistics. Under the null hypothesis each of the log-rank statistics (3.1) are asymptotically independent so, for instance, a chi-squared statistic can be computed based on $\sum \bar{U}_s^j(l,m)^2$ where the sum is over all l, m where a direct $l \to m$ is possible and $l \in \mathcal{R}_j \bigcap \mathcal{R}_{j^c}$. A test based on the maximum of the individual statistics would also be possible, but is not pursued here.

3.2 Global test

The local test statistics can be computed at any time within the follow up period leading to a family of statistics. As such, s is now considered to vary across some interval $s \in [t_0, t_{\max}] \subset [0, \tau]$, which represents the period of follow-up in which the Markov property is to be tested. Then, under a null hypothesis H_0 : $\alpha_{lm}(t|X(s) = j) = \alpha_{lm}(t|X(s) \neq j)$ for all $t_0 \leq s \leq t_{\max}, s \leq t \leq \tau$, the process $\{\bar{U}_s^{(j)}(l,m), s \in [t_0, t_{\max}]\}$ converges to a zero mean Gaussian process with a covariance function that can be consistently estimated. This motivates the development of 'global' tests based on summary statistics of $\{\bar{U}_s^{(j)}(l,m), s \in [t_0, t_{\max}]\}$, for instance $\sup_{s\in[t_0,t_{\max}]} |\bar{U}_s^{(j)}(l,m)|$, $\int_{t_0}^{t_{\max}} |\bar{U}_s^{(j)}(l,m)|ds$ or more generally $\int_{t_0}^{t_{\max}} w(s)|\bar{U}_s^{(j)}(l,m)|ds$, for some weight function $w(s)$, which test the dependence of $\alpha_{lm}(t)$ on occupancy in state j for the $l \to m$ transition over the full range of follow-up. Note also that for a fixed j, the processes $\bar{U}_s^{(j)}(l,m)$ and $\bar{U}_s^{(j)}(l',m')$, where $l \neq l'$ or $m \neq m'$, are asymptotically independent. Thus the individual supremum or integrated statistics may be combined to provide an overall test of dependence of the transition intensities on occupation in state j at previous times. Moreover, this also provides an overall test for consistency of the AJ estimator of $\mathbf{P}_j(s,t)$. Note that the local log-rank test statistic in (3.1) will be undefined or unstable in time periods in which very few patients are either in, or not in, the qualifying state j. As such, $[t_0, t_{\max}]$ should be chosen to avoid periods, typically at the beginning or end of follow-up, where this may occur. Alternatively, weights $w(s)$ can be chosen to downweight such periods. This issue is discussed further in Section 3.7.

3.3 Null distribution of the statistic

To determine the null distribution of the proposed statistics, consider first the log-rank statistic for a specific transition $(l \to m)$, from a fixed starting time s and state of interest j. In order to simplify the notation, for the remainder of this section, unless otherwise specified, dependence on (l,m) of the various quantities

will be suppressed, such that $N_i(t)$ be the counting process for $(l \to m)$ transitions for subject i by time t, including the possibility of multiple transitions of the same type and $Y_i(t) = Y_{il}(t)$ be the at risk indicator of $l \to m$ for subject i. The asymptotic covariance between the statistics (3.1) at time points s and s' where $s \leq s'$ is given by

$$\mathrm{Cov}(U_s^{(j)}, U_{s'}^{(j)}) = \sum_i \int_{s'}^{\tau} Y_i(t) \left\{ \delta_i^{(j)}(s) - \frac{\sum_k \delta_k^{(j)}(s) Y_k(t)}{\sum_k Y_k(t)} \right\} \left\{ \delta_i^{(j)}(s') - \frac{\sum_k \delta_k^{(j)}(s') Y_k(t)}{\sum_k Y_k(t)} \right\} d\hat{A}(t), \quad (3.2)$$

where $\hat{A}(t) = \int_0^t \sum_k dN_k(u) / \sum_k Y_k(u)$ is the Nelson-Aalen estimator.

At each unique transition time $t_{(k)} \geq s$, the number of patients at risk can be summarized as $n_k = \sum_i Y_i(t_{(k)})$ comprised of $n_k = n_{k11} + n_{k01} + n_{k10} + n_{k00}$ where $n_{k11} = \sum_i Y_i(t_{(k)}) \delta_i^{(j)}(s) \delta_i^{(j)}(s')$, $n_{k01} = \sum_i Y_i(t_{(k)})(1 - \delta_i^{(j)}(s)) \delta_i^{(j)}(s')$, $n_{k10} = \sum_i Y_i(t_{(k)}) \delta_i^{(j)}(s)(1 - \delta_i^{(j)}(s'))$ and $n_{k00} = \sum_i Y_i(t_{(k)})(1 - \delta_i^{(j)}(s))(1 - \delta_i^{(j)}(s'))$. An estimate of the covariance for a given dataset is then (for $s \leq s'$)

$$\widehat{\mathrm{Cov}}(U_s^{(j)}, U_{s'}^{(j)}) = \sum_{k: t_{(k)} \geq s'} \frac{n_{k11} n_{k00} - n_{k01} n_{k10}}{n_k^2}.$$

When $s = s'$ the formula reduces to the standard asymptotic approximation for the variance term in the log-rank statistic

$$\widehat{\mathrm{Var}}(U_s^{(j)}) = \sum_{k: t_{(k)} \geq s} \frac{n_{k11}(n_k - n_{k11})}{n_k^2}. \quad (3.3)$$

Under the null hypothesis, $\{\bar{U}_s^{(j)}, s \in [t_0, t_{\max}]\}$ converges to a zero mean Gaussian process with covariance function $\mathrm{Cov}(U_s^{(j)}, U_{s'}^{(j)}) / \sqrt{\mathrm{Var}(U_s^{(j)}) \mathrm{Var}(U_{s'}^{(j)})}$.

3.4 Computation of the null distribution

The value of $\bar{U}_s^{(j)}$ for the $(l \to m)$ transition may change at any time of entry into state l, and any time of exit or censoring from state l. As such, it is computationally expensive to determine $\bar{U}_s^{(j)}$ at all times. Instead, a grid of times, $s_1, \ldots, s_L$, at which the statistic is to be computed can be pre-specified and the statistic taken as $\max_l \bar{U}_{s_l}^{(j)}$ or $L^{-1} \sum_l |\bar{U}_{s_l}^{(j)}|$. The asymptotic null distribution can then be approximated by simulating from the multivariate normal distribution that arises if the Gaussian process is evaluated at the grid points $s_1, \ldots, s_L$.

However, a better small sample approximation can be obtained by using a wild bootstrap (Lin *and others*, 1993) in which the increments of the martingale processes are multiplied by i.i.d. random variables, $G_{ih}, i = 1, \ldots, n,\ h = 1, \ldots, N_i(\tau)$, with mean 0 and variance 1. Beyersmann *and others* (2013) found better small sample results using centered Poisson random variables, rather than standard normal weights.

Hence for an original realization of (3.1)

$$u_s^{(j)} = \sum_{i=1}^{n} \sum_{h=N_i(s)+1}^{N_i(\tau)} \left(\delta_i^{(j)}(s) - \frac{\sum_k \delta_k^{(j)}(s) Y_k(t_{ih})}{\sum_k Y_k(t_{ih})} \right),$$

the wild bootstrap version would be

$$u_s^{(j)*} = \sum_{i=1}^{n} \sum_{h=N_i(s)+1}^{N_i(\tau)} \left(\delta_i^{(j)}(s) - \frac{\sum_k \delta_k^{(j)}(s) Y_k(t_{ih})}{\sum_k Y_k(t_{ih})} \right) G_{ih},$$

where t_{ih} corresponds to the hth $l \to m$ transition time for subject i.

3.5 Extension to Cox-Markov models

An analogous test is possible for Cox-Markov multi-state models where the process is Markov conditional on observed covariates that have a proportional effect on the transition intensities, as defined in (2.1). The log-rank test of a difference between two groups can be derived through the score test of a Cox proportional hazards model with a single binary covariate. Hence the test statistic $U_s^{(j)}$ given in (3.1) is the score statistic of the covariate $\delta_i^{(j)}(s)$ using data at times $t > s$. In the presence of additional covariates $\mathbf{z}$ with associated coefficients $\boldsymbol{\beta}_0$, the score statistic becomes

$$U^{(j)}(s; \boldsymbol{\beta}_0) = \sum_i \int_s^\tau \left\{ \delta_i^{(j)}(s) - \bar{\Delta}_s^{(j)}(t; \boldsymbol{\beta}_0) \right\} dN_i(t),$$

where

$$\bar{\Delta}_s^{(j)}(t; \boldsymbol{\beta}_0) = \frac{\sum_k \delta_k^{(j)}(s) Y_k(t) \exp(\boldsymbol{\beta}_0 \mathbf{z}_k)}{\sum_k Y_k(t) \exp(\boldsymbol{\beta}_0 \mathbf{z}_k)},$$

which provides a statistic for testing the hypothesis $H_0 : \alpha_{lm}(t; \mathbf{z}|X(s) = j) = \alpha_{lm}(t; \mathbf{z}|X(s) \neq j)$ for $t \geq s$. In practice, $\boldsymbol{\beta}_0$ will be unknown and instead $U^{(j)}(s; \hat{\boldsymbol{\beta}})$ is used, where $\hat{\boldsymbol{\beta}}$ is the maximum partial likelihood estimate of $\boldsymbol{\beta}_0$.

Let $M_i(t) = N_i(t) - \int_0^t Y_i(u) \exp(\boldsymbol{\beta}_0 \mathbf{z}_i) \alpha_0(u) du$ denote the martingale process associated with patient i. Using arguments similar to those in Lin *and others* (1993), $n^{-1/2} U^{(j)}(s; \hat{\boldsymbol{\beta}})$ is asymptotically equivalent to the process $n^{-1/2} \tilde{U}^{(j)}(s; \boldsymbol{\beta}_0)$, where

$$\begin{aligned} \tilde{U}^{(j)}(s; \boldsymbol{\beta}_0) = & \sum_{i=1}^{n} \int_s^\tau \left\{ \delta_i^{(j)}(s) - \tilde{\Delta}_s^{(j)}(t; \boldsymbol{\beta}_0) \right\} dM_i(t) \\ & - \sum_{k=1}^{n} \int_s^\tau \left\{ \delta_k^{(j)}(s) - \tilde{\Delta}_s^{(j)}(t; \boldsymbol{\beta}_0) \right\} \left\{ \mathbf{z}_k - \tilde{Z}(t; \boldsymbol{\beta}_0) \right\} Y_k(t) \exp(\boldsymbol{\beta}_0 \mathbf{z}_k) \alpha_0(t) dt \\ & \times \mathcal{I}(\boldsymbol{\beta}_0)^{-1} \sum_{i=1}^{n} \int_0^\tau \left\{ \mathbf{z}_i - \tilde{Z}(t; \boldsymbol{\beta}_0) \right\} dM_i(t), \end{aligned}$$

where $\tilde{\Delta}_s^{(j)}(t;\boldsymbol{\beta}_0)$ is the limit of $\bar{\Delta}_s^{(j)}(t;\boldsymbol{\beta}_0)$, $\tilde{Z}(t;\boldsymbol{\beta}_0)$ is the limit of $\bar{Z}(t;\boldsymbol{\beta}_0) = \frac{\sum_{k=1}^n Y_k(t)\mathbf{z}_k \exp(\boldsymbol{\beta}_0\mathbf{z}_k)}{\sum_{k=1}^n Y_k(t)\exp(\boldsymbol{\beta}_0\mathbf{z}_k)}$ and $\mathcal{I}(\boldsymbol{\beta}_0)$ is the Fisher information associated with $\boldsymbol{\beta}$ using all event time data.

The asymptotic equivalence of the two processes allows a wild bootstrap technique to be applied to approximate the distribution of summaries of the process. Again using i.i.d. random variables, G_{ih}, a wild bootstrap replication of the process can be obtained via:

$$\hat{U}^{(j)}(s;\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}\sum_{h=N_i(s)+1}^{N_i(\tau)} \left\{\delta_i^{(j)}(s) - \bar{\Delta}_s^{(j)}(t_{ih};\hat{\boldsymbol{\beta}})\right\} G_{ih} - \mathbf{h}^{(j)}(s,\hat{\boldsymbol{\beta}})\sum_{i=1}^{n}\sum_{h=1}^{N_i(\tau)} \left\{\mathbf{z}_i - \bar{Z}(t_{ih};\hat{\boldsymbol{\beta}})\right\} G_{ih}, \quad (3.4)$$

where

$$\mathbf{h}^{(j)}(s,\hat{\boldsymbol{\beta}}) = \sum_{k=1}^{n}\int_s^\tau \left\{\delta_k^{(j)}(s) - \bar{\Delta}_s^{(j)}(t,\hat{\boldsymbol{\beta}})\right\}\left\{\mathbf{z}_k - \bar{Z}(t;\hat{\boldsymbol{\beta}})\right\} Y_k(t)\exp(\hat{\boldsymbol{\beta}}\mathbf{z}_k) d\hat{A}(t;\hat{\boldsymbol{\beta}}) \times \mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}$$

and $\hat{A}(t;\hat{\boldsymbol{\beta}}) = \int_0^t \frac{\sum_k dN_k(u)}{\sum_k Y_k(u)\exp(\hat{\boldsymbol{\beta}}\mathbf{z}_k)}$ is the Breslow estimator of the baseline cumulative intensity.

To produce a test statistic directly analogous to that of the case with no covariates, we work with an approximately standardized process $\bar{U}^{(j)}(s;\hat{\boldsymbol{\beta}}) = U^{(j)}(s;\hat{\boldsymbol{\beta}})/\sqrt{V^{(j)}(s;\hat{\boldsymbol{\beta}})}$ where

$$V^{(j)}(s;\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}\int_0^\tau \left(\left\{\delta_i^{(j)}(s) - \bar{\Delta}_s^{(j)}(t,\hat{\boldsymbol{\beta}})\right\} I(t>s) - \mathbf{h}^{(j)}(s,\hat{\boldsymbol{\beta}})\left\{\mathbf{z}_i - \bar{Z}(t;\hat{\boldsymbol{\beta}})\right\}\right)^2 Y_i(t)\exp(\hat{\boldsymbol{\beta}}\mathbf{z}_i) d\hat{A}(t;\hat{\boldsymbol{\beta}}).$$

Provided there are separate estimates $\hat{\boldsymbol{\beta}}_{lm}$ for each transition intensity ($l \to m$), the test statistics may be considered to be asymptotically independent under the null. The tests then proceed in a similar way as in the case of no covariates. An extension to the case where there are common β that act upon more than one transition intensity (Putter *and others*, 2007) is possible, but is beyond the scope of the current paper.

3.6 Transition specific statistic

When a multi-state model has several states, there will often be more than two qualifying states, j, which are relevant to a specific transition intensity, e.g. $l \to m$. If interest lies in determining whether the Markov property holds generally, then multiple indicators, $\delta_i(s)^{(j)} = I(X_i(s) = j)$ may be defined for each $j \in \mathcal{R}_l^*$, where $\mathcal{R}_l^*$ is the set of states from which l is reachable. For each j we can define a score statistic, $U^{(j)}(s,\hat{\boldsymbol{\beta}})$, as defined before. Let $\mathbf{U}(s;\hat{\boldsymbol{\beta}})$ be a vector whose jth element corresponds to $U^{(j)}(s,\hat{\boldsymbol{\beta}})$, for $j \in \mathcal{R}_l^*$, and let $\boldsymbol{\Psi}$ be the square singular matrix whose (j,j') element corresponds to the covariance between $U^{(j)}(s,\hat{\boldsymbol{\beta}})$ and $U^{(j')}(s,\hat{\boldsymbol{\beta}})$. Asymptotically under the null hypothesis, $\mathbf{U}(s;\hat{\boldsymbol{\beta}})$ has mean zero and hence

$K(s;\hat{\boldsymbol{\beta}}) = \mathbf{U}'_{(r)}\boldsymbol{\Psi}^{-1}_{(r,r)}\mathbf{U}_{(r)}$ is asymptotically $\chi^2_{r^*-1}$, where $r^* = |\mathcal{R}^*_l|$, $\mathbf{U}_{(r)}$ denotes the vector $\mathbf{U}(s;\hat{\boldsymbol{\beta}})$ with the rth element removed and $\boldsymbol{\Psi}_{(r,r)}$ represents $\boldsymbol{\Psi}$ with the rth row and column removed. The element r can be chosen arbitrarily from the state space $\{1,\ldots,R\}$. Section S2 of the Supplementary Material gives the form of $\boldsymbol{\Psi}$.

A global test for the $l \to m$ transition can be constructed by considering an appropriate summary, such as the mean or maximum, of $K(s;\hat{\boldsymbol{\beta}})$ along $(t_0, t_{\max}]$. The null distribution of the global test statistic may again be approximated by a wild bootstrap applying the same set of multipliers $\{G_{ih}\}$ to each of the score functions as defined in (3.4). The statistics, $K^{(lm)}(s,\hat{\boldsymbol{\beta}}_{lm})$, relating to different transitions $l \to m$ are again asymptotically independent if the process is Markov and separate $\hat{\boldsymbol{\beta}}_{lm}$'s are estimated for each transition intensity. An overall test for the whole process can be defined by combining the individual summaries in some way, for instance through the maximum, mean or weighted mean.

3.7 Choice of summaries and weights

As noted above, the processes $\bar{U}^{(j)}(s;\hat{\boldsymbol{\beta}})$ and $K(s;\hat{\boldsymbol{\beta}})$ may be summarized through absolute mean, absolute maximum statistics or weighted absolute mean statistics. For unweighted mean or maximum statistics it will generally be necessary to truncate the range of times s, for which $\bar{U}^{(j)}(s;\hat{\boldsymbol{\beta}})$ or $K(s;\hat{\boldsymbol{\beta}})$ is calculated, to exclude times in which few patients are at risk or where most or all patients are in one qualifying state. Equally, for a weighted statistic such time points should also be down-weighted. In the Supplementary Material, Section S6, it is argued that

$$r_j(s) = \frac{\sqrt{d_j(s)n_j^{(1)}(s)n_j^{(0)}(s)}}{n_j^{(1)}(s) + n_j^{(0)}(s)}, \tag{3.5}$$

where $n_j^{(1)}(s) = \sum_i \delta_i^{(j)}(s)Y_i(s)$, $n_j^{(0)}(s) = \sum_i (1-\delta_i^{(j)}(s))Y_i(s)$ and $d_j(s) = \bar{N}(\tau) - \bar{N}(s)$, can be used either as the weight function or as a criterion for deciding at which time point to truncate the test range. When used as weight function in $\int_{t_0}^{\tau} r_j(s)|\bar{U}^{(j)}(s;\hat{\boldsymbol{\beta}})(l,m)|ds$ or in the weighted chi-square overall test statistic, we propose to first standardize $r_j(s)$ such that $\int_{t_0}^{\tau} r_j(s) = \tau - t_0$, even though it will have no effect on the resulting test statistics.

Let $\bar{U}_{lm}^{(j)}$ and $\bar{K}_{lm}$ denote the preferred summary statistic in each case. An overall statistic for state j may be constructed by taking either a maximum value statistic $U^* = \max_{l,m} \bar{U}_{lm}^{(j)}$ or a mean value statistic $U^* = N_T^{-1}\sum_{l,m}\bar{U}_{lm}^{(j)}$, where N_T is the number of statistics to be combined. However, some transitions may be more represented in the data, so it may be better to consider a weighted mean value, such as $U^* = \frac{\sum_{l,m} n_{lm}\bar{U}_{lm}^{(j)}}{\sum_{l,m} n_{lm}}$, where $n_{l,m}$ is the total number of $l \to m$ transitions observed within the dataset.

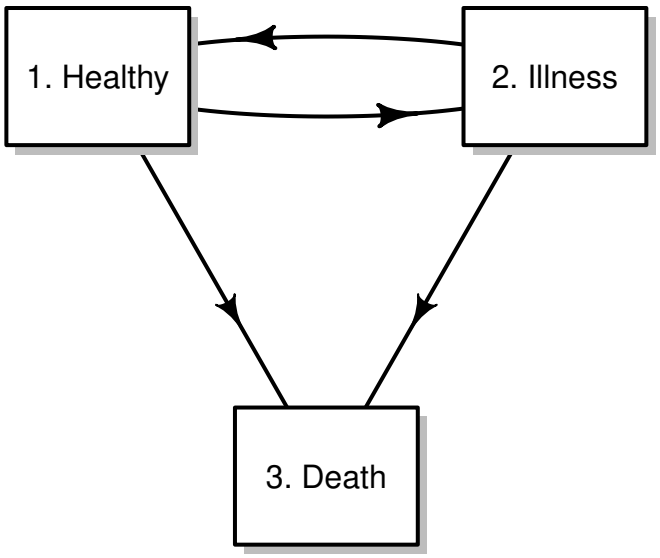

Figure 1: State diagram of the reversible illness-death model

The same principles could also be applied when combining the transition specific statistics $K(s;\hat{\boldsymbol{\beta}})$. However, the size of the set $\mathcal{R}_l^*$ may vary for different l. As such, some method of p-value combination is perhaps the most convenient approach. For instance, a weighted Pearson combination rule of the form $K^* = -\sum_{l,m} w_{lm} \log p_{lm}^*$ where p_{lm}^* is the p-value associated with the summary of the $l \to m$ and $w_{l,m}$ is some associated weight. Sensible choices of weights would either be constant weights, or else to take $w_{lm} = n_{lm}$.

4 Simulation study

In this section a simulation study is conducted to evaluate the performance of the existing and proposed statistics for realistic sample sizes, under a range of scenarios.The simulations concentrate on the case of a three-state illness-death model, shown in Figure 1 where recovery is possible (state 1 = healthy, state 2 = illness, state 3 = death). The process has four distinct transition intensities $(\alpha_{12}, \alpha_{13}, \alpha_{21}, \alpha_{23})$. All transitions can occur at time t if $X(s) \in \{1, 2\}$; as such for each of the transitions, the associated log-rank process $\bar{U}^{(2)}(s)$, which takes $j = 2$ as the qualifying state, is sufficient to test the Markov property. Six main scenarios are considered. The first four are (a) a Markov scenario to allow assessment of control of Type I error, (b) a homogeneous semi-Markov (or clock reset) model, (c) a shared frailty model where the frailty affects all transitions, and (d) a shared frailty model where the frailty effect acts differently for the backward 'illness' to 'healthy' transition.

Scenario (e) is designed to illustrate a scenario where the Cox-based test will perform poorly and assumes it is the duration in the previous state, rather than duration in the current state, that drives the transitions out of each state. However, the intensities are chosen in such a way that the current duration in the state has no marginal effect (see Section S3 in the Supplementary Material for details). In scenario (f) the non-Markovian behaviour is governed by the state occupied at time $t = 3$, with all intensities changing if $X(3) = 1$, which is a situation where the log-rank based test will perform optimally.

In addition, a Cox-Markov model scenario is generated where there is a single time fixed covariate with a standard normal distribution that is assumed to be known to affect the $1 \to 2$ and $2 \to 1$ transitions only, with $\beta_{12} = 1$ and $\beta_{21} = -1$. This Cox-Markov case is carried out for $n = 100$ and $n = 500$. In Section S4 of the Supplementary Material additional simulations are performed on a three-state recurrent process, similar to that of the application in Section 5 to investigate the performance of the transition specific statistic.

For each scenario, we consider $n = 100$ and $n = 500$ subjects and a 50% probability of starting in state 1 as opposed to state 2 at time 0. The data are assumed to be continuously observed up to independent random right-censoring at time C_i where $C_i \sim U(5, 25)$. In all cases except scenario (f), the baseline intensities are taken to be Weibull, with shape parameters $\alpha = (1.2, 0.8, 1.4, 1.0)$ and rate parameters $\lambda = (0.18, 0.03, 0.25, 0.1)$, for transitions $1 \to 2$, $1 \to 3$, $2 \to 1$, and $2 \to 3$, respectively. These intensities are shown in Figure S3 in the Supplementary Material. For the shared frailty models, the frailties u_i are gamma distributed with mean 1 and variance 0.5. For scenario (c) each intensity is multiplied by the same factor u_i. For scenario (d) each intensity is multiplied by u_i, except the $2 \to 1$ intensity, which is multiplied by $1/\sqrt{u_i}$. For the pathological non-Markov case, scenario (f), the shape parameters are unchanged with respect to the previous scenarios, but rate parameters are $\lambda = (0.18, 0.03, 0.25, 0.1)$ for $t < 3$ and for $t \geq 3$ if $X(3) \neq 1$, while $\lambda = (0.27, 0.06, 0.15, 0.2)$ if $t \geq 3$ and $X(3) = 1$. For the non-Markov scenarios with $n = 500$, the parameters governing the non-Markov behaviour are scaled by $1/\sqrt{5}$ to make the power approximately comparable to the $n = 100$ case.

In all cases, the global statistic for $\bar{U}^{(2)}(s)$ is computed on a fixed grid of points $\mathbf{s}$ at increments of 0.1. Six ways of generating an overall statistic are considered: the summary statistics, $\bar{U}^2_{lm}$, are taken to be either a maximum absolute value or mean absolute value based on a fixed time range $[0.5, 10]$, or weighted using (3.5) using all follow-up times, and in each case the overall statistic is constructed either as the mean of the summary statistics or as a weighted mean. Note that, since patients start in state 1 or state 2 with probability 0.5, there is not a problem with imbalance between groups at early follow-up times in this case. At $t = 10$, on average around 45 patients are still at risk and there will be patients at risk in both state 1

and state 2 in the vast majority of simulated datasets.

4.1 Simulation results

To indicate the degree of bias in the transition probability estimates arising in the non-Markov scenarios, Figure S2 in the Supplementary material presents the average estimates of transition probabilities using the standard Aalen-Johansen estimator (AJ) and using the landmark Aalen-Johansen estimator, based on three different starting times ($s = 3, 6, 9$).

Figure 2 shows the empirical distribution of p-values over 5000 simulation replicates for the proposed log-rank based test, as well as the Cox proportional hazards based test using the time of entry into the current state as a covariate, and the Commenges-Andersen test, for each of the six scenarios. Table 1 presents the empirical power for the different tests based on a nominal 5% Type I error, under the six scenarios considered. For scenario (d), the Commenges-Andersen test statistic using just the forward transitions is also presented (denoted Sub C-A).

In the Markov case, scenario (a), both variants of the wild bootstrap test statistic give reasonably close to nominal 5% Type I error. However, the statistic based on absolute maximum value has a distribution of p-values further from uniform. When $n = 100$, the test statistics are conservative, with empirical type I errors ranging from 3.3% to 4.5% for a nominal 5% rate. This behaviour disappears when $n = 500$. The null behaviour of the Cox-Markov variant of the test behaves similarly, with the test being slightly more conservative than the case without covariates when $n = 100$, but with good performance for $n = 500$.

For the semi-Markov case, scenario (b), the Cox test has substantially better power than either log-rank based statistics. The test based on the mean absolute value has slightly higher power than the absolute maximum. The Commenges-Andersen test has essentially no power to detect a difference, which is perhaps not surprising since each transition within a patient is still independent within the model.

For scenario (c), the shared frailty with common multiplicative frailty effect on all transitions, the Cox test does substantially better than either of the log-rank function based tests. In this case, the absolute maximum version is more powerful than the absolute mean based test. The Commenges-Andersen test greatly outperforms the others, having more than 99% power.

When the shared frailty has differing effects for the forward compared to backward transitions, scenario (d), the respective performances of the Cox test and log-rank function based tests are more similar, with the weighted absolute mean variant outperforming the Cox test. The absolute mean version performs better than the absolute maximum. Despite a shared frailty, the Commenges-Andersen test based on all transitions

Table 1: Empirical power of tests for nominal 5% Type I error under different scenarios. Mean, Max and WMean refer to absolute mean, absolute maximum (within the range [0,10]) and weighted mean statistics (using the full follow-up time) to compute each $\bar{U}_{lm}^{(j)}$. Equal combination refers to computing U^* by a simple mean of the individual statistics. Weighted combination refers to using a weighted mean based on the total number of observation transitions of each type. *Power is 0.7012 if the subset version of C-A is applied. **Power is 0.9634 if the subset version of C-A is applied. Deviations from Markov have been scaled in the $n = 500$ case to produce approximately comparable power.

		Equal combination			Weighted combination				
	Scenario	Mean	Max	WMean	Mean	Max	WMean	Cox	C-A
(a)	Markov ($n = 100$)	0.0326	0.0346	0.0344	0.0428	0.0446	0.0478	0.0550	0.0494
	Markov ($n = 500$)	0.0500	0.0496	0.0480	0.0534	0.0540	0.0552	0.0514	0.0506
	Cox-Markov ($n = 100$)	0.0356	0.0334	0.0258	0.0382	0.0398	0.0260	0.0508	0.0486
	Cox-Markov ($n = 500$)	0.0488	0.0480	0.0414	0.0464	0.0504	0.0366	0.0494	0.0516
(b)	Semi-Markov ($n = 100$)	0.1498	0.1344	0.1490	0.1756	0.1492	0.1826	0.5846	0.0598
	Semi-Markov ($n = 500$)	0.1856	0.1398	0.1994	0.2062	0.1542	0.2212	0.6164	0.0690
(c)	Shared frailty1 ($n = 100$)	0.4454	0.3602	0.4618	0.5546	0.4972	0.5824	0.8650	0.9972
	Shared frailty1 ($n = 500$)	0.5148	0.4066	0.5628	0.6142	0.5078	0.6508	0.9754	1.0000
(d)	Shared frailty2 ($n = 100$)	0.4844	0.3674	0.4912	0.5804	0.4706	0.5926	0.5200	0.2426*
	Shared frailty2 ($n = 500$)	0.7658	0.6294	0.8104	0.8158	0.7038	0.8534	0.7382	0.4176**
(e)	Duration ($n = 100$)	0.1978	0.1234	0.2060	0.1854	0.1168	0.1892	0.0794	0.0508
	Duration ($n = 500$)	0.2112	0.1664	0.2270	0.1710	0.1390	0.1816	0.0668	0.0502
(f)	Pathological ($n = 100$)	0.4970	0.6390	0.4216	0.5884	0.7294	0.5368	0.0688	0.0568
	Pathological ($n = 500$)	0.6786	0.8104	0.5938	0.7138	0.8494	0.6556	0.0750	0.0510

has substantially lower power than either the Cox test or the log-rank based tests, implying that it is strongly affected by contrary frailty effects. However, if the Commenges-Andersen test is applied only to the subset of transitions for which the positive dependence applies (i.e. $1 \to 2, 1 \to 3$ and $2 \to 3$), then a test more powerful than the log-rank based test is again obtained.

For scenario (e), the previous duration dependent case, the log-rank based tests perform substantially better than the Cox test, which has very modest power, and the Commenges-Andersen test, which has essentially a uniform distribution of p-values. In this case, the power of the absolute mean statistic is superior to the absolute maximum.

The pathological non-Markov case, scenario (f), is the setting where the proposed log-rank test would be expected to perform best. The extent to which it outperforms the Cox test is nevertheless impressive. Neither the Cox nor the Commenges-Andersen tests have any discernible power to detect the non-Markov behaviour, whereas both log-rank based tests have at least 50% power. The maximum absolute based test performs somewhat better than using the mean.

In all non-Markov scenarios except (f), using a weighted absolute mean statistic based weighting by (3.5) resulted in a more powerful test than using equal weights over the range $s \in [0.5, 10]$. However, the weighted test remains conservative for the Cox-Markov model even when $n = 500$.

Figure S4 in the Supplementary Material gives the empirical distribution of p-values using transition count weighted combinations. In the null Markov case, the resulting p-values have distributions closer to a uniform distribution. For scenarios (c), (d) and (f), weighting by number of transitions improves the power of the log-rank based tests, whereas for (b) and (e) the power is the same or slightly worse. Broadly similar results were seen for scenarios (b)-(f) using $n = 500$ rather than $n = 100$. Figures S5 and S6 give the corresponding plots of distribution of p-values using weighted and unweighted combinations, respectively.

5 Application

We illustrate our methods using data on sleeping behaviour collected at the Max-Planck Institute for Psychiatry in Munich as part of a larger study on sleep withdrawal. The data were also analysed in Yassouridis *and others* (1999) and Kneib and Hennerfeind (2008). The sleep process is recorded for 70 subjects every 30 seconds by electroencephalographic (EEG) measurements from the moment of first falling asleep until finally waking up, which are afterwards classified into three states: 1. Awake, 2. non-REM sleep, 3. REM sleep. Transitions between any pair of distinct states are possible, leading to a total of six possible transi-

tions. The multi-state process is highly dynamic: a total of 4637 transitions were observed, the majority from Awake to non-REM sleep (1660) and from non-REM sleep to Awake (1368). Figures S7 and S8 in the Supplementary Material show the non-parametric estimates of the cumulative intensities of all transitions, and the Aalen-Johansen estimates of the state occupation probabilities, respectively.

It is expected that inter-individual differences in transition intensities (frailties) will play a dominating role in this multi-state process. This suspicion is confirmed by the Commenges-Andersen test ($p \approx 5 \cdot 10^{-45}$), and by the log-rank tests developed in this paper also yielding extremely significant deviations from Markovianity (the corresponding log-rank process plots are given in Figures S10 and S11 in the Supplementary Material). We therefore sought for evidence for departures from Markovianity beyond these inter-individual differences in transition intensities. A rigorous analysis of this question would be outside the scope of this paper; since the objective is to illustrate the log-rank tests developed in this paper, we took the following pragmatic approach. Following the ideas of Kneib and Hennerfeind (2008) we started by fitting a frailty model with independent frailties for each subject by transition combination. As in Kneib and Hennerfeind (2008) the frailties were all taken to be independent; probably this is sub-optimal, but it is outside the scope of this paper to attempt to fit correlated frailty models. The empirical Bayes estimates for each subject by transition combination were then calculated and their logarithm was used as offsets in the log-rank test statistics. In simulations presented in Section S4 of the Supplementary Material, substituting the estimated frailties is observed to lead to a conservative test. This is because each patient's martingale process will satisfy $M_i(\tau) \approx 0$ and hence $U_s^{(j)}$ will tend to be less variable than if the true frailty were known.

The correlations between the transition-specific frailty estimates were generally positive; the correlation between the log empirical Bayes estimates for non-REM to REM and REM to non-REM was 0.57, between non-REM to REM and REM to Awake was 0.40; between Awake to REM and REM to Awake was also 0.40; the rest of the correlations were between -0.28 and 0.37. Figure S9 in the Supplementary Material shows histograms and scatterplots of the transition-specific empirical Bayes frailty estimates.

When discussing the results of the log-rank tests we will concentrate on the transition from Awake to non-REM sleep and the transition back from non-REM sleep to Awake. Figure 3 shows in solid black the log-rank test statistic process $\{\bar{U}_s^{(j)}(l,m), s \in [t_0, \tau]\}$, for $(l,m) = (1,2)$, i.e. for the transition from Awake to non-REM sleep, for an equally spaced grid of time points s from $t_0 = 1$ minute to $\tau = 8$ hours. The trace is shown for each of the qualifying states j. In grey are the first 50 of 1000 wild bootstrap traces; the inner black dotted lines are the pointwise 2.5% and 97.5% quantiles of the wild bootstrap traces, while the outer black dotted lines are plus and minus the 95% quantile of the wild bootstrap replicates of the supremum

Table 2: Summary log-rank test statistics (associated p-values) for two selected transitions

	Awake → non-REM sleep							
	Awake		non-REM		REM		Overall	
Unweighted	1.50	(0.000)	1.07	(0.004)	1.01	(0.002)	5.23	(0.000)
Proposed								
weighted	1.48	(0.000)	1.04	(0.007)	1.06	(0.013)	4.16	(0.001)
Supremum	6.34	(0.003)	5.26	(0.008)	4.59	(0.025)	40.23	(0.003)
	non-REM sleep → Awake							
	Awake		non-REM		REM		Overall	
Unweighted	1.13	(0.003)	0.91	(0.127)	0.80	(0.172)	3.48	(0.001)
Proposed								
weighted	1.05	(0.021)	0.84	(0.315)	0.85	(0.280)	2.28	(0.132)
Supremum	6.64	(0.001)	6.75	(0.002)	3.27	(0.254)	53.91	(0.001)

of $|\bar{U}_s^{(j)}|$. We consider the following three summary statistics of these processes: the unweighted average $\int_{t_0}^{\tau} |\bar{U}_s^{(j)}(l,m)|ds$, the proposed weighted average $\int_{t_0}^{\tau} w(s)|\bar{U}_s^{(j)}(l,m)|ds$, with $w(s)$ defined as in (3.5), and the supremum statistic $\sup_{s\in[t_0,\tau]} |\bar{U}_s^{(j)}(l,m)|$. Table 2 shows the results of the log-rank tests with these three summary statistics, together with the overall chi-square test statistic. The general picture arising from these tests is quite clear; for the transition from Awake to non-REM sleep, irrespective of the summary test statistic and of the qualifying state the Markov assumption after having accounted for the frailties is rejected. The only exception is for the weighted summary statistic and qualifying state 2, non-REM sleep.

Interestingly, in contrast to the simulation study except for the pathological case, the results for the weighted summary statistic are less significant than for the unweighted ones. This is even clearer for the backward transition from non-REM sleep to Awake. The log-rank test statistics in Figure S12 in the Supplementary Material show that the clearest departures from Markovianity appear towards the end of the sleeping period. The supremum test statistics confirm the statistical significance of these departures from Markovianity for $s > 7$, except for REM sleep as qualifying state. The average based test statistics generally fail to pick up this behaviour, with only the unweighted test statistic for qualifying state 1, Awake, reaching statistical significance. Again the weighted average test statistics are systematically less significant than the unweighted average. An explanation can be found from a plot of the proposed weights for the transition

non-REM to Awake, and qualifying states Awake and non-REM, shown in Figure S13 of the Supplementary Material, after s=6 for both qualifying states $w(s)$ steadily decreases towards 0, and the weights are near zero when the log-rank test statistics in Figure S12 are largest.

6 Discussion

When applying multi-state models in practice, it is important to check the Markov assumption. Many violations of the Markov assumption can be anticipated: a transition intensity can depend on the sojourn time or frailty can be present. For irreversible illness-death models a general test of the Markov assumption is available (Rodríguez-Girondo and de Uña-Álvarez, 2012), but otherwise only *ad hoc* tests for certain aspects of non-Markovianity exist. This paper has provided approaches of testing the most common departures of Markovianity, the Cox modelling approach and the stratified Commenges-Andersen test, and has proposed a general class of log-rank type test statistics that have adequate power against a range of alternatives.

In the absence of prior knowledge on the likely violation from a Markov process, we recommend to use the transition-specific chi-square test statistic of Section 3.6 for each transition, where weights for each qualifying state, as defined in Section 3.7, are used. An adjustment for multiple testing based on the number of transitions is appropriate. Since the original Markov model is defined in terms of, broadly separate, models for the transition intensities, the tests can inform a researcher on which transitions may require more complicated modelling. In the event of a significant test result for a given intensity, the log-rank process plots can give some indication of the possible nature of the problem. Figure S15 in the Supplementary Material gives the mean and standard deviation of $\bar{U}$ statistics over time for a selection of the simulation scenarios. If the overall statistic for transition $l \to m$ is significant and the log-rank statistics given occupancy in state l have the same sign for all, or most, of the time range, it might indicate a persistent semi-Markov type effect similar to the assumption of a Cox model with time of entry into the state as a covariate. In contrast if there is only a persistent effect for part of the time scale, as is the case for Awake $\to$ non-REM given Awake in the sleep example (Figure 3), then this indicates a more localized effect e.g. that if a subject is awake at some point after 6 hours since sleep onset then it will have an effect on their ability to fall back to sleep whereas no such tendency is apparent in the first part of the night. Frailty effects are harder to diagnose as they may exhibit patterns to the mean similar to a semi-Markov effect. However, the presence of frailty may have the effect of increasing the general variability of the log-rank process, as can be seen in the respective log-rank processes for the sleep dataset, with and without accounting for frailties (Figures 3 and S10).

We expect testing on a transition-specific basis to have adequate power to detect the most common departures of Markovianity; our application showed the supremum version to be more significant than the weighted summary statistic, but it should be noted that this is after having accounted for the more dominating – and more common – departure from Markovianity due to frailty. The test can be used as a pre-test to determine whether to use Aalen-Johansen (AJ), valid under the Markov assumption, or more robust approaches like the landmark Aalen-Johansen (LMAJ) estimator, to calculate transition probabilities.

It should be noted in this context that even when the Markov assumption is not satisfied, the AJ estimator may have smaller mean squared error than the LMAJ estimator, as shown in simulation studies (Putter and Spitoni, 2018). It is the familiar bias-variance trade-off, where for smaller sample size variance tends to dominate – in favour of AJ – and for larger sample size bias tends to dominate – in favour of LMAJ. Using this work as a pre-test fits in nicely within this framework; larger sample size will have more power to detect violations of the Markov assumption, suggesting to use robust methods. More work is needed to study how this works out in practice.

7 Software

The R code for implementing the proposed test on a simulated dataset similar to the data analysed in Section 5 is available at `https://github.com/andrewtitman/MarkovTest`.

Acknowledgement

Thomas Kneib is gratefully acknowledged for making available the sleeping behaviour data.

Supplementary Material

Supplementary material is available online at `http://biostatistics.oxfordjournals.org`.

References

Aalen, O. O. and Johansen, S. (1978). An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scandinavian Journal of Statistics. Theory and Applications* **5**, 141–150.

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.

Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research* **11**, 91–115.

Balan, T. A. and Putter, H. (2019). frailtyem: An R package for estimating semiparametric shared frailty models. *Journal of Statistical Software* **90**, 1–29.

Beyersmann, J., Pauly, M. and Di Termini, S. (2013). Weak Convergence of the Wild Bootstrap for the Aalen–Johansen Estimator of the Cumulative Incidence Function of a Competing Risk. *Scandinavian Journal of Statistics* **40**, 387–402.

Chiou, S. H., Qian, J., Mormino, E. and Betensky, R. A. (2018). Permutation tests for general dependent truncation. *Computational Statistics & Data Analysis* **128**, 308–324.

Commenges, D. and Andersen, P.K. (1995). Score test of homogeneity for survival data. *Lifetime Data Analysis* **1**, 145–156.

Cook, R.J. and Lawless, J.F. (2018). *Multistate models for the analysis of life history data*. Boca Raton: CRC Press.

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.

Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics* **42**, 855–865.

Kneib, Thomas and Hennerfeind, Andrea. (2008). Bayesian semi parametric multi-state models. *Statistical Modelling* **8**, 169–198.

Lin, D. Y., Wei, L. J. and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–572.

Meira-Machado, L., 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.

Meira-Machado, L., de Uña-Álvarez, J., 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.

Putter, H., Fiocco, M. and Geskus, R.B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine* **26**, 2389–2403.

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**, 2081–2092.

Putter, H. and van Houwelingen, H. C. (2015). Frailties in multi-state models: Are they identifiable? Do we need them? *Statistical Methods in Medical Research* **24**, 675–692.

Rodríguez-Girondo, M. and de Uña-Álvarez, J. (2012). A nonparametric test for Markovianity in the illness-death model. *Statistics in Medicine* **31**, 4416–4427.

Titman, A. C. (2015). Transition Probability Estimates for Non-Markov Multi-State Models. *Biometrics* **71**, 1034–1041.

Yassouridis, Alexander, Steiger, Axel, Klinger, Artur and Fahrmeir, Ludwig. (1999). Modelling and exploring human sleep with event history analysis. *Journal of Sleep Research* **8**, 25–36.

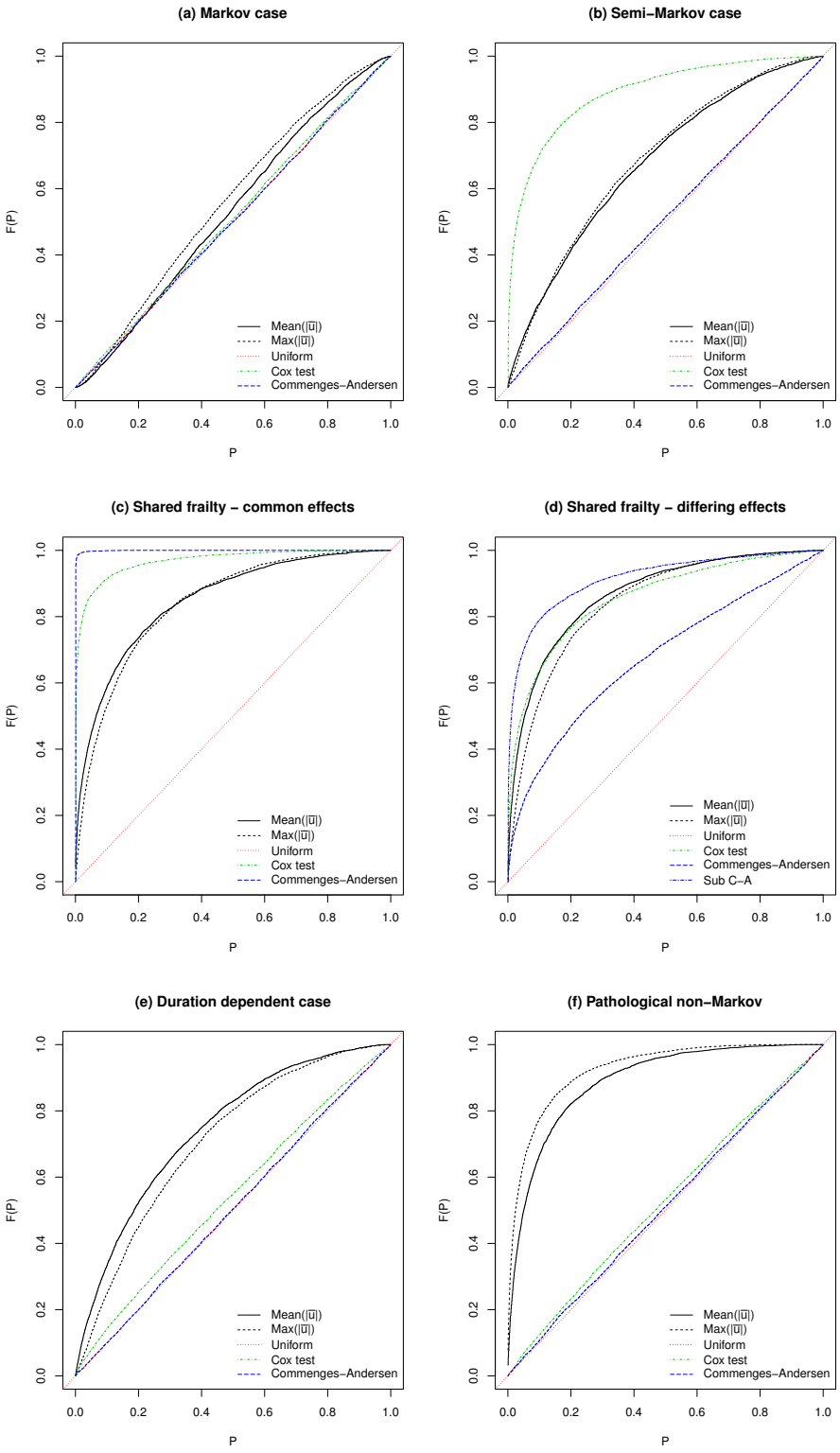

Figure 2: Empirical distribution of p-values for the six scenarios using unweighted combinations of the log-rank statistics when $n = 100$.

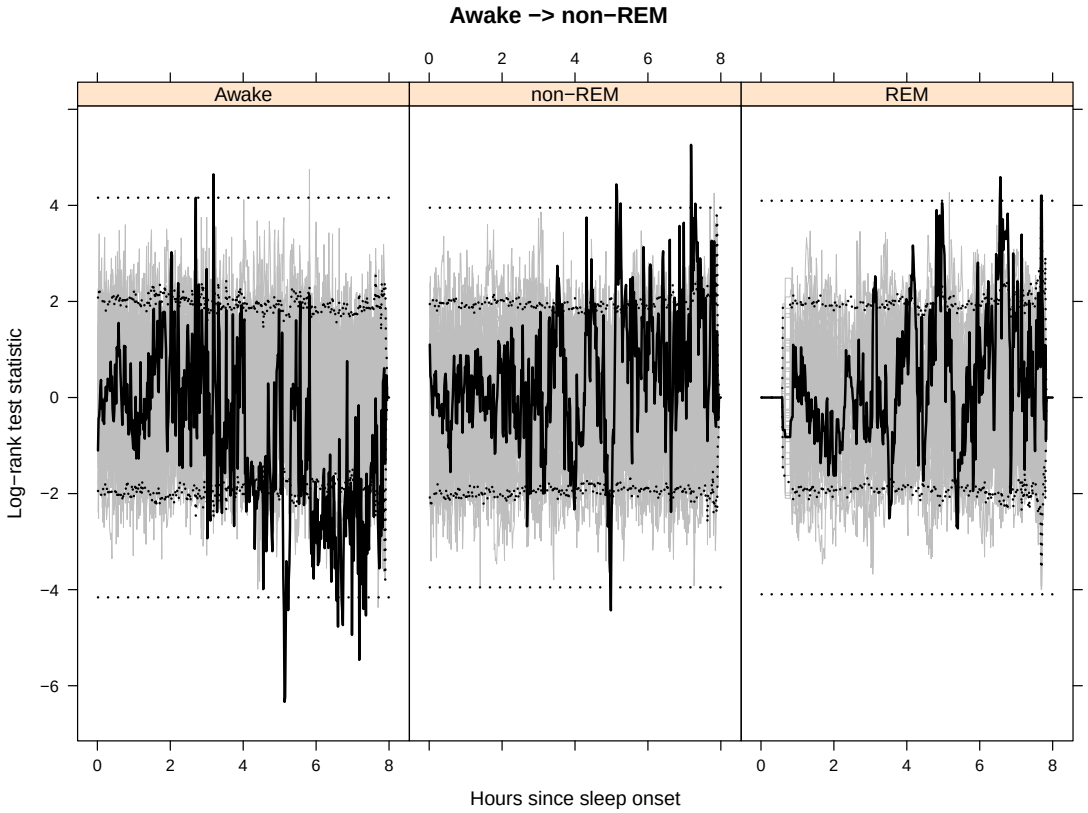

Figure 3: Log-rank process for the transition from Awake to non-REM sleep