Dynamic inference for non-Markov transition probabilities under random right censoring

Dennis Dobler\textsuperscript{1} | Andrew Titman\textsuperscript{2}

\textsuperscript{1}Department of Mathematics, Faculty of Science, Vrije Universiteit Amsterdam
\textsuperscript{2}Department of Mathematics & Statistics, Lancaster University

Correspondence
Dennis Dobler, Department of Mathematics, Faculty of Sciences, Vrije Universiteit Amsterdam, De Boelelaan 1111, 1081 HV Amsterdam, The Netherlands.
Email: d.dobler@vu.nl

Funding information
Deutsche Forschungsgemeinschaft, BE-4500 1-2

Abstract
The main contribution of this article is the verification of weak convergence of a general non-Markov (NM) state transition probability estimator by Titman, which has not yet been done for any other general NM estimator. A similar theorem is shown for the bootstrap, yielding resampling-based inference methods for statistical functionals. Formulas of the involved covariance functions are presented in detail. Particular applications include the conditional expected length of stay in a specific state, given occupation of another state in the past, and the construction of time-simultaneous confidence bands for the transition probabilities. The expected lengths of stay in a two-sample liver cirrhosis dataset are compared and confidence intervals for their difference are constructed. With borderline significance and in comparison to the placebo group, the treatment group has an elevated expected length of stay in the healthy state given an earlier disease state occupation. In contrast, the Aalen-Johansen (AJ) estimator-based confidence interval, which relies on a Markov assumption, leads to a drastically different conclusion. Also, graphical illustrations of confidence bands for the transition probabilities demonstrate the biasedness of the AJ estimator in this scenario.
data example. The reliability of these results is assessed in a simulation study.

**KEYWORDS**
confidence bands, Markov assumption, multistate model, restricted conditional expected length of stay, right censoring, weak convergence

1 | INTRODUCTION

Transition probabilities are essential quantities in survival analytic examinations of multistate models. Under independent right censoring, for instance, the Aalen-Johansen (AJ) estimator assesses these probabilities optimally in nonparametric Markovian multistate models (cf. Aalen, 1978 and sect. IV.4 in Andersen, Borgan, Gill, & Keiding, 1993 for more details). The Markov assumption is crucial for the validity of the AJ estimator. However, in real world applications, it may be unrealistic to postulate such a structure: given the patient’s present state, future developments are often not independent of a severe past illness history. There have been several attempts to circumvent the Markov assumption: State occupation probabilities in general models have been estimated by Datta and Satten (2001, 2002) and Glidden (2002) using the AJ estimator. Gunnes, Borgan, and Aalen (2007) investigated the robustness of the Datta-Satten and AJ estimates of the state occupation probabilities in the presence of covariates that affect both the transition intensities and censoring hazard. Pepe, Longton, and Thornquist (1991) and Pepe (1991) used a combination of two Kaplan-Meier estimators to assess the relevant transition probability in an illness-death model with recovery. In illness-death models without recovery, Meira-Machado, de Uña-Álvarez, and Cadarso-Suárez (2006) used Kaplan-Meier-based techniques and consecutive latent transition times to estimate transition probabilities. More efficient variants of these estimators have been developed by de Uña-Álvarez and Meira-Machado (2015). de Uña-Álvarez and Mandel (2018) considered estimation in general progressive multistate models for right-censored and left-truncated observations. A Kendall’s $\tau$-based test of the Markov assumption in the progressive illness-death model was derived by Rodríguez-Girondo and Uña-Álvarez (2012). Meira-Machado, de Uña-Álvarez, and Datta (2015) developed a nonparametric estimator of transition probabilities conditional on covariates in the illness-death model without recovery, and the censoring times may depend on the covariates. Azarang, Scheike, and de Uña-Álvarez (2017) proposed methodology for direct regression of transition probabilities, without a Markov requirement, using estimating equations based on the binomial response of occupancy in a particular state. The recent article by Meira-Machado and Sestelo (2019) gives a good overview of estimation methods available in illness-death models without recovery.

A competing risks-based estimator was proposed by Allignol, Beyersmann, Gerds, and Latouche (2014); they and Meira-Machado et al. (2006) rely on the assumption that the support of the censoring distribution is included in the support of the survival distribution, which Allignol et al. extended to the usual random censorship model using artificial censoring. Allignol et al. (2014) constructed their estimator based on a subsampling technique, which only includes those individuals which were observed to be in the initial state of interest at the initial time of interest. Other estimation procedures were also built on this subsampling idea (cf. de Uña-Álvarez & Meira-Machado, 2015; Putter & Spitoni, 2018; Titman, 2015); Putter and Spitoni called it
landmarking. The estimators by Titman (2015) and Putter and Spitoni (2018) are applicable in general non-Markov (NM) models. Both use different transition probability decompositions and utilizations of the AJ estimator and they do not rely on strong assumptions on the censoring times, just the common random right censoring assumption. However, weak convergence properties of these estimators as elements of càdlàg function spaces have not been analyzed yet.

In the present article, we focus on the NM state transition probability estimator by Titman (2015) and its weak convergence for the following reasons: the estimator has a simple, but intuitive structure and it allows estimation of general transition probabilities between sets of states rather than single states. Finally, the bootstrap is shown to correctly reproduce the weak limit distribution. Throughout the article, let \((\Omega, \mathcal{A}, P)\) be the underlying probability space. To introduce the Titman (2015) estimator, we consider an independently right-censored multistate model with \(R + 1 \in \mathbb{N}\) different states. That is, the data consist of independent copies \(X_1, \ldots, X_n\) of a multistate process, which are indexed by time \(t \geq s\) and which occupy the states \(0, 1, \ldots, R\). The right censoring is modeled by independent random variables \(C_1, \ldots, C_n\) after which further observation of the corresponding multistate processes is no longer possible. Let \(I\) and \(J\) be subsets of \(\{0, 1, \ldots, R\}\) for which the transition probabilities

\[
P_{IJ}(s, t) = P(X(t) \in J|X(s) \in I),
\]

shall be estimated. Here and below, irrelevant subscripts are removed for notational convenience. Introduce the absorbing subset of states \(A_J \subset J\) that implies sure occupancy of \(J\) at all later points in time, and the absorbing subset of states \(R_J \subset \{0, 1, \ldots, R\}\setminus J = \overline{J}\) which prevents occupancy of \(J\) at all later points of time. Individual competing risks processes, \(Z_i, i = 1, \ldots, n\), are used in an intermediate estimation step. Here, \(Z_i(t) = 0, 1, 2\), respectively, imply occupation of \(A_J \cup R_J, A_J,\) and \(R_J\) at time \(t\). Let \(Y_i(t), i = 1, \ldots, n\) denote the left-continuous at-risk indicators of the competing risks processes and let \(\delta_i, i = 1, \ldots, n\) be the corresponding censoring indicators, that is, \(\delta_i = 1\) if the transition of \(Z_i\) is eventually observed and \(\delta_i = 0\) if a censoring comes first. The actually observable data are summarized in \(\chi = \{X_i \cdot Y_i, Y_i, \delta_i : i = 1, \ldots, n\}\). Hence, the processes \(X_i\) (and \(Z_i\)) are observable only until censoring occurs. The transition probability of interest can be decomposed into three probabilities that are estimable based on \(\chi\):

\[
P_{IJ}(s, t) = F_1(t) + F_0(t)p_{J|I}(t)
\]

\[
= P(Z(t) = 1|X(s) \in I) + P(Z(t) = 0|X(s) \in I)P(X(t) \in J|Z(t) = 0, X(s) \in I);
\]

(Titman, 2015). The first two components in this decomposition are estimated by the standard Kaplan-Meier estimator \(\hat{F}_0\) for the survival function and the AJ estimator \(\hat{F}_1\) for the cumulative incidence function of the subgroup of those individuals, which were observed to occupy state \(I\) at time \(s\). The remaining conditional probability is estimated using the empirical proportion

\[
\hat{p}_{J|I}(t) = \frac{\hat{p}_{IJ}(t)}{\hat{p}_I(t)} = \frac{n^{-1} \sum_{\ell=1}^n 1\{X_\ell(s) \in I, X_\ell(t) \in J, Y_\ell(t) = 1\}}{n^{-1} \sum_{\ell=1}^n 1\{X_\ell(s) \in I, Y_\ell(t) = 1\}},
\]

where \(1\{\cdot\}\) denotes the indicator function. Consequently, the overall transition probability estimator by Titman (2015) is given as

\[
\hat{P}_{IJ}(s, t) = \hat{F}_1(t) + \hat{F}_0(t)\hat{p}_{J|I}(t).
\]
This article shall establish the weak convergence of $\sqrt{n}(\hat{P}_{IJ}(s, \cdot) - P_{IJ}(s, \cdot))$ and a bootstrap variant thereof as $n \to \infty$. Finally, implications to confidence interval construction for the conditional expected length of stay and to time-simultaneous confidence bands for the transition probabilities are demonstrated. The proofs of the central limit theorems and a detailed derivation of the asymptotic covariance function are given in the Supplementary Material provided online.

2 | MAIN RESULTS

We focus on the Titman (2015) estimator during $[s, \tau]$ where $\tau \in (s, \infty)$ is a terminal time, such that $F_0(\tau) > 0$ and $P(C > \tau) > 0$. The large sample properties of $\hat{P}_{IJ}(s, \cdot)$ are derived from central limit theorems for its individual components. To obtain such for the fraction process $\hat{p}_{JI}$, we throughout assume the existence of bounded one- and two-step hazard functions for transitions into $J$ and $\overline{J}$, respectively, which are approached uniformly:

\[
\lim_{\delta \downarrow 0} \sup_{t-t_1=\delta} \left| \frac{P(X(t_1) \notin J, X(t) \in J)}{t-t_1} - \lambda_J(t) \right| = 0, \quad (1)
\]

\[
\lim_{\delta \downarrow 0} \sup_{t-t_1=\delta} \left| \frac{P(X(t_1) \notin J, X(t) \in J)}{t-t_1} - \lambda_{\overline{J}}(t) \right| = 0, \quad (2)
\]

\[
\lim_{\delta \downarrow 0} \sup_{t-t_1=\delta} \left| \frac{P(X(t_2) \notin J, X(t) \in J)}{t_2-t} - \lambda_{\overline{J}|J}(t_2|t_1) \right| = 0, \quad (3)
\]

\[
\lim_{\delta \downarrow 0} \sup_{t-t_1=\delta} \left| \frac{P(X(t_2) \notin J, X(t) \in J)}{t_2-t} - \lambda_{\overline{J}|\overline{J}}(t_2|t_1) \right| = 0. \quad (4)
\]

Here, $t_1, t, t_2 \in (s, \tau]$. Conditions (1) and (2) impose slightly more regularity than just the existence of the hazard rates $\lambda_J, \lambda_{\overline{J}}$, and the same is true for the two-step hazard functions in (3) and (4). Loosely speaking, the latter two conditions imply that multiple state changes into or out of $J$ do not happen in a too fluctuating way. In the special case when $X$ is Markov, conditions (3) and (4) reduce to (1) and (2), respectively. Another realistic example, in which the above four conditions are met, is given in the data generating process for our simulations in Section 4.

From now on, we slightly modify the original definition of $\hat{p}_{JI}$ by using the right-continuous versions of the at-risk processes, indicated by a + sign in the argument. This leads to the choice

\[
\hat{p}_{JI}(t) = \frac{n^{-1} \sum_{\ell=1}^{n} 1\{X_\ell(s) \in I, X_\ell(t) \in J, Y_\ell(t+) = 1\}}{n^{-1} \sum_{\ell=1}^{n} 1\{X_\ell(s) \in I, Y_\ell(t+) = 1\}}.
\]

The reason for this choice is that, using the same notation as before, the resulting estimator $\hat{P}_{IJ}(s, \cdot)$ can then be considered as a random element of the Skorokhod space $D[s, \tau]$ of right-continuous functions with existing left-sided limits. The probabilistic consequences of this
slightly modified estimator $\hat{P}_{IJ}$ are negligible because all event times are continuously distributed, so that a censoring coincides with an event with probability 0. After these preparations, we come to our main central limit theorem.

**Theorem 1.** Under conditions (1) to (4) and as $n \to \infty$, the following convergence holds in distribution on $D[s, \tau]$:

$$\sqrt{n}(\hat{P}_{IJ}(s, \cdot) - P_{IJ}(s, \cdot)) \to U,$$

where $U$ is a zero-mean Gaussian process.

The asymptotic covariance function $(r, t) \mapsto \Gamma_{IJ}(r, t)$ of $U$ is given in Section 3 in the Supplementary Material. To be able to use the result of Theorem 1 for time-simultaneous inference on $P_{IJ}(s, \cdot)$ or simply for improving the accurateness of $\hat{P}_{IJ}(s, \cdot)$-based inference methods with a pivotal limit distribution, the classical bootstrap seems to be a useful choice of a resampling method. To this end, a bootstrap sample $\chi^*$ is obtained by randomly drawing triples $(X_i, Y_i, \delta_i)$ $n$ times with replacement from the original sample $\chi$. This bootstrap sample is used to recalculate the transition probability estimator, resulting in

$$P^*_I(s, t) = F^*_1(t) + F^*_0(t)p^*_J(t), \quad t \in [s, \tau].$$

For this bootstrap counterpart, a central limit theorem similar to the main Theorem 1 holds:

**Theorem 2.** Under conditions (1) to (4), as $n \to \infty$, and given $\chi$, the following conditional convergence in distribution on $D[s, \tau]$ holds in probability:

$$\sqrt{n}(P^*_I(s, \cdot) - \hat{P}_{IJ}(s, \cdot)) \to U,$$

where $U$ is the same Gaussian process as in Theorem 1.

That is, the bootstrap succeeds in correctly recovering the limit distribution of the transition probability estimate $\hat{P}_{IJ}(s, \cdot)$, enabling an approximation of its finite sample distribution.

### 3 | INFERENTIAL PROCEEDURES

#### 3.1 | Restricted conditional expected length of stay

A practical application of both main theorems is a two-sample comparison of expected lengths of stay in the state set of interest, $\mathcal{J}$, given occupation of a fixed state set $\mathcal{I}$ at time $s \in [0, \tau]$. Considering temporarily the above one-sample setup, the restricted conditional expected length of stay is

$$\bar{e}_{IJ}(s, \tau) = \mathbb{E} \left[ \int_s^\tau \mathbf{1}(X(u) \in \mathcal{J}) \, du \mid X(s) \in \mathcal{I} \right]$$

$$= \int_s^\tau P(X(u) \in \mathcal{J} \mid X(s) \in \mathcal{I}) \, du.$$

Such quantities could be of clinical interest in assessing the effect of complications to surgical interventions (Silber et al., 1999). Grand and Putter (2016) considered a pseudo-observation
regression treatment in the context of estimating expected years of health or disability conditional on the state occupied at a given age. An R package for deriving the change in lengths of stay, that relies on the AJ estimator, is described in Wangler, Beyersmann, and Schumacher (2006). For further literature on lengths of stay within multistate models, see the articles cited therein. In the special case of a simple survival model, the only expected length of stay of interest is the mean residual lifetime; Irwin (1949) proposed to consider such expected lifetimes which are restricted to a fixed end point in time. Here, the Markov assumption shall not be imposed and estimation of the length of stay is based on the Titman (2015) estimator.

Sample-specific quantities are furnished with a superscript $(\cdot)$, $e = 1, 2$. For instance, $\hat{P}_{IJ} (s, t)$ is the first sample’s transition probability estimate to switch into some state of $I$ at time $t$ given that a state of $I$ is occupied at time $s$. The sample sizes $n_1$ and $n_2$ may be different. We would like to test the null hypothesis

$$H_{J|I} : \epsilon_{IJ}^{(1)} (s, \tau) \geq \epsilon_{IJ}^{(2)} (s, \tau) \quad \text{versus} \quad K_{J|I} : \epsilon_{IJ}^{(1)} (s, \tau) < \epsilon_{IJ}^{(2)} (s, \tau).$$

If $J$ is a favourable set of states, then the rejection of $H_{J|I}$ is an indication of the superiority of the second treatment over the first. A reasonable estimator of $\epsilon_{IJ}^{(e)} (s, \tau)$ is $\hat{\epsilon}_{IJ}^{(e)} (s, \tau) = \int_s^\tau \hat{P}_{IJ}^{(e)} (s, u) du$, $e = 1, 2$. Note that, even under the boundary $\partial H_{J|I}$ : $\epsilon_{IJ}^{(1)} (s, \tau) = \epsilon_{IJ}^{(2)} (s, \tau)$, the underlying transition probabilities $P_{IJ}^{(1)}$ and $P_{IJ}^{(2)}$ are allowed to differ. Hence, we do not rely on the restrictive null hypothesis $\tilde{H}$ : $P_{IJ}^{(1)} \equiv P_{IJ}^{(2)}$ but rather on the hypothesis $H_{J|I}$ of actual interest.

**Theorem 3.** Under conditions (1) to (4) and if $0 < \lim \inf n_1 / n_2 \leq \lim \sup n_1 / n_2 < \infty$ as $n \to \infty$, the following convergence in distribution holds:

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\epsilon_{IJ}^{(1)} (s, \tau) - \hat{\epsilon}_{IJ}^{(2)} (s, \tau)}{\hat{\sigma}_{IJ}} \to N(0, 1).$$

Here, $\hat{\sigma}_{IJ}^2$ is any consistent estimator, for example, found via plug-in, of the asymptotic variance

$$\sigma_{IJ}^2 = (1 - \lambda) \int_s^\tau \int_s^\tau \Gamma_{IJ}^{(1)} (u, v) dv du + \lambda \int_s^\tau \int_s^\tau \Gamma_{IJ}^{(2)} (u, v) dv du,$$

along subsequences of $n_1 / (n_1 + n_2) \to \lambda \in (0, 1)$.

Replacing all estimators in the previous theorem by their bootstrap counterparts, where both underlying bootstrap samples $\chi^{(1)*}$ and $\chi^{(2)*}$ are obtained by $n_1$ and $n_2$ times independently and separately drawing with replacement from $\chi^{(1)}$ and $\chi^{(2)}$, respectively, a similar convergence in conditional distribution given $\chi^{(1)} \cup \chi^{(2)}$ holds in probability. This result may be used for bootstrap-based and asymptotically exact confidence intervals for the difference in the restricted conditional expected lengths of stay, $\epsilon_{IJ}^{(1)} (s, \tau) - \epsilon_{IJ}^{(2)} (s, \tau)$; see Sections 4 and 5 below.

### 3.2 | Simultaneous confidence bands

Even though the bootstrap procedure is useful but not strictly necessary for inference on the restricted expected length of stay, it plays an essential role in time-simultaneous inference on the transition probabilities $t \mapsto P_{IJ} (s, t)$. This is due to the unknown stochastic behaviour of the
limit process $U$ which, again in the context of creating confidence bands, is also the reason for
the inevitability of resampling procedures for AJ estimators even if the Markov assumption is
fulfilled; see Bluhmki et al. (2018) for theoretical justifications and their practical performance.

For the derivation of reasonable time-simultaneous $1 - \alpha \in (0,1)$ confidence bands for
$P_{t,J}(s, \cdot)$ on a time interval $[t_1, t_2] \subset [s, \tau]$ based on $\hat{P}_{t,J}(s, \cdot)$ in combination with the bootstrap, we
focus on the process

$$B_n : t \mapsto \sqrt{n} \cdot w(t) \cdot (\phi(\hat{P}_{t,J}(s, t)) - \phi(P_{t,J}(s, t))), \quad t \in [t_1, t_2],$$

where a transformation $\phi$ will ensure bands within the probability bounds 0% and 100%, and a
suitable multiplicative weight function $w$ will stabilize the bands for small samples. In particular,
we use one of the log-log transformations $\phi_1(p) = \log(-\log(p))$ or $\phi_2(p) = \log(-\log(1-p))$ (Lin,
1997), depending on whether we expect $P_{t,J}(s, t_1)$ to be closer to 1 or 0. As weight function, for
transformation $k \in \{1, 2\}$, we choose $w_{1k}(t) = [\phi_k'(\hat{P}_{t,J}(s, t)) \cdot \hat{\sigma}_{t,J}(t)]^{-1},$

$$w_{21}(t) = -\left(\phi_1'(\hat{P}_{t,J}(s, t)) \cdot \hat{P}_{t,J}(s, t) \cdot \left[1 + \frac{\hat{\sigma}_{t,J}^2(t)}{\hat{P}_{t,J}^2(s, t)}\right]\right)^{-1}$$

$$= -\log(\hat{P}_{t,J}(s, t)) / \left[1 + \frac{\hat{\sigma}_{t,J}^2(t)}{\hat{P}_{t,J}^2(s, t)}\right],$$

or

$$w_{22}(t) = \left(\phi_2'(\hat{P}_{t,J}(s, t)) \cdot (1 - \hat{P}_{t,J}(s, t)) \cdot \left[1 + \frac{\hat{\sigma}_{t,J}^2(t)}{(1 - \hat{P}_{t,J}(s, t))^2}\right]\right)^{-1}$$

$$= -\log(1 - \hat{P}_{t,J}(s, t)) / \left[1 + \frac{\hat{\sigma}_{t,J}^2(t)}{(1 - \hat{P}_{t,J}(s, t))^2}\right].$$

With these choices, the resulting confidence bands will correspond to the classical equal precision
(EP) bands for $w = w_{1k}$ or the Hall-Wellner (H-W) bands for $w = w_{2k}$ if $J$ is an absorbing
subset of states (cf. p. 265ff in Andersen et al., 1993 for the survival case and $k = 1$ or Lin, 1997 in
the presence of competing risks and $k = 2$).

The confidence bands are found by solving for $P_{t,J}(s, t)$ in the probability
$P(\sup_{t \in [t_1, t_2]} |B_n(t)| \leq q_{1-\alpha}) = 1 - \alpha$. Here, the value of $q_{1-\alpha}$ is approximated by the
$(1 - \alpha)$-quantile $q_{1-\alpha}^*$ of the conditional distribution of the supremum of the bootstrap version $B_n^*(t) = \sqrt{n} \cdot \hat{w}^*(t) \cdot (\phi(\hat{P}_{t,J}^*(s, t)) - \phi(\hat{P}_{t,J}(s, t)))$ of $B_n$ given the data, where in the definition of $\hat{w}^*$ all estimators $\hat{P}_{t,J}(s, t)$ and $\hat{\sigma}_{t,J}(t)$ are
replaced with their bootstrap counterparts $\hat{P}_{t,J}^*(s, t)$ and $\hat{\sigma}_{t,J}^*(t)$.

Finally, the resulting $1 - \alpha$ confidence bands for $t \mapsto P_{t,J}(s, t)$ are as follows:

$$CB_{1k} = [\phi_1^{-1}(\phi_1(\hat{P}_{t,J}(s, t)) \mp q_{1-\alpha}^*/(\sqrt{n}w_{1k}(t)))]_{t_1 \leq t \leq t_2}$$

$$= [\hat{P}_{t,J}(s, t) \exp(\mp q_{1-\alpha}^*/(\sqrt{n}w_{1k}(t)))]_{t_1 \leq t \leq t_2}$$

and

$$CB_{2k} = [\phi_2^{-1}(\phi_2(\hat{P}_{t,J}(s, t)) \mp q_{1-\alpha}^*/(\sqrt{n}w_{2k}(t)))]_{t_1 \leq t \leq t_2}$$

$$= [1 - (1 - \hat{P}_{t,J}(s, t)) \exp(\mp q_{1-\alpha}^*/(\sqrt{n}w_{2k}(t)))]_{t_1 \leq t \leq t_2}$$

The performance of these confidence bands as well as the confidence band obtained from not
transforming or weighting the transition probability estimator are assessed in a simulation study
in the subsequent sections.
SIMULATION STUDY

4.1 Restricted conditional expected length of stay

To assess the performance of the conditional length of stay estimators for finite sample sizes, a small simulation study is conducted. Data are simulated from a pathological NM model in which the subsequent dynamics of the process depend on the state occupied at \( t = 4 \). A three-state illness-death model with recovery is used where the state 0 means “healthy”, 1 “ill”, and 2 “dead”. The transition intensities are \( \lambda_{02}(t) = 0.02, \lambda_{10}(t) = 0.3, \lambda_{12}(t) = 0.1 \), while

\[
\lambda_{01}(t) = \begin{cases} 
0.3 & \text{if } X(4) = 1, t > 4 \\
0.6 & \text{otherwise}.
\end{cases}
\]  

(7)

In Section 7 of the Supplementary Material, we prove that the conditions (1) to (4) are satisfied in the present setup. Subjects are independently right-censored via an exponential distribution with rate 0.04. We focus on the restricted expected length of stay in the health state conditional on an earlier illness at time \( s = 5 \), that is, \( I = \{1\} \) and \( J = \{0\} \). For each simulated dataset, \( e_{10}(5, \tau) \) is estimated and 95% confidence intervals are constructed via three methods: a Wald interval using an estimator \( \hat{\sigma}^2_{10} \) of the asymptotic variance \( \sigma^2_{10} \), which is found by plugging in canonical estimators, a classical bootstrap, and a bootstrap-t procedure using the bootstrap-version of \( \hat{\sigma}^2_{10} \) to studentize the bootstrap estimator. For each bootstrap, \( B = 1000 \) samples are generated. In the present three-state model, the Titman (2015) transition probability estimator reduces to \( \hat{P}_{10}(5, t) = \hat{P}_{I,J}(5, t) = \hat{F}_0(t)\hat{p}_{J|I}(t) = \hat{F}_0(t)\hat{p}_{0|1}(t) \) because the healthy state is nonabsorbing. As a competing method, AJ estimator-based Wald-type confidence intervals are constructed for the same expected length of stay. Standard arguments yield that the asymptotic covariance function of the normalized AJ estimator for the \( 1 \rightarrow 0 \) transition is

\[
(r, t) \mapsto \sum_{j=0}^{2} \sum_{k \neq j} \int_s^t \int_s^t P^2_{1j}(s, u)(P_{k0}(u, t) - P_{j0}(u, t))(P_{k0}(u, r) - P_{j0}(u, r)) \frac{\alpha_{jk}(u)}{y_j(u)} du;
\]

(sect. IV.4.2 in Andersen et al., 1993 for the corresponding particular variance formula). The expected length of stay asymptotic variance again results from a double integral of this covariance function with integration range \( r, t \in (s, \tau] \).

We consider several different scenarios for the total sample size, \( n = 50, 100, 150, \) and 200. Note that the NM estimator only uses the subgroup of subjects satisfying \( C_i > 5, X_i(t) = 1 \), which corresponds to only 0.367\( n \) subjects, on average. For each of the scenarios, 10 000 datasets are generated. As a result, the empirical coverage percentages have approximate Monte-Carlo standard error of around 0.2.

The simulation results are shown in Table 1. Both the AJ and NM estimator have substantial bias when \( n = 50 \) leading to under-coverage of all the nominal 95% confidence intervals. For larger \( n \), both of the bootstrap confidence intervals and the Wald-based interval for the NM estimator give close to nominal coverage. The coverage of the AJ-based Wald confidence interval deteriorates with increasing \( n \) due to the estimator being inherently biased.
**TABLE 1** Empirical bias and coverage of nominal 95% confidence intervals for the restricted conditional expected length of stay based on simulated datasets

| $n$  | Bias | Coverage % | Wald (AJ) | Wald (NM) | Naive bootstrap | Bootstrap-$t$ |
|------|------|------------|-----------|-----------|----------------|--------------|
| 50   | −0.231 | 0.181      | 92.09     | 91.01     | 94.23          | 90.85        |
| 100  | −0.268 | 0.085      | 89.95     | 94.54     | 94.08          | 94.43        |
| 150  | −0.279 | 0.063      | 88.14     | 95.24     | 94.21          | 95.00        |
| 200  | −0.297 | 0.033      | 86.34     | 95.56     | 94.54          | 95.43        |

Abbreviations: AJ, Aalen-Johansen; NM, non-Markov.

### 4.2 Simultaneous confidence bands

To investigate the coverage probabilities of simultaneous confidence bands introduced in Section 3.2, datasets are simulated under the same NM model as above, and considering total sample sizes of $n = 50, 100, 150, 200, 300,$ and $400$ patients. Ninety-five percent simultaneous confidence bands are constructed for $p_{21}(5, t)$ for $t \in (6, 7]$. For the NM estimator, in addition to the H-W and EP bands using the transformation $\phi_2(p)$, a naive confidence band based on a constant weight function, $w(t) \equiv 1$, and an identity transformation, $\phi(p) = p$, is also constructed. In addition, for the AJ estimates, EP bands based on $\phi_2(p)$, are constructed both via the Efron bootstrap and also via the wild bootstrap using the R code which has been made available in the supplement to Bluhmki et al. (2018). In all cases, for each bootstrap $B = 1000$ samples are generated.

Table 2 gives the empirical coverage probabilities of the simultaneous confidence bands based on 5000 simulated datasets. For the bands based on the NM estimator, the H-W and EP bands tend to over-cover quite markedly for small sample sizes, while the naive bands under-cover. Adequate coverage is achieved for the H-W and EP bands by $n = 200$. Since the AJ estimator is biased for this scenario, we would expect the coverage of the EP AJ bands to deteriorate with increasing sample size. However, it appears this is counteracted by the estimated quantiles, $q_{1-a}$ increasing with $n$. As a consequence, the coverage initially grows with increasing $n$ before beginning to decrease. There is more marked under-coverage for the Efron than wild bootstrap constructed EP AJ bands. This is not surprising given the fact that Efron’s bootstrap does not succeed in reproducing an

**TABLE 2** Empirical coverage of nominal 95% simultaneous confidence bands for $p_{21}(5, t)$ for $t \in (6, 7]$

| $n$  | H-W | NM | EP | Naive | EP (WB) | EP (Efron) |
|------|-----|----|----|-------|--------|-----------|
| 50   | 0.9996 | 0.9972 | 0.9226 | 0.9154 | 0.9784 |
| 100  | 0.9896 | 0.9846 | 0.9238 | 0.9188 | 0.9170 |
| 150  | 0.9604 | 0.9542 | 0.9298 | 0.9230 | 0.9060 |
| 200  | 0.9496 | 0.9496 | 0.9420 | 0.9124 | 0.9068 |
| 300  | 0.9476 | 0.9466 | 0.9408 | 0.9154 | 0.8998 |
| 400  | 0.9440 | 0.9440 | 0.9396 | 0.9102 | 0.8974 |

Abbreviations: AJ, Aalen-Johansen; EP, equal precision; H-W, Hall-Wellner; NM, non-Markov; WB, wild bootstrap.
appropriate randomness nature of the trajectories—even in the Markov case. The reason for this is that, in the case of continuous event times, the drawn trajectory is uniquely determined by the first event time.

5 | APPLICATION TO THE LIVER CIRRHOSIS DATASET

As an illustrative example, we consider the liver cirrhosis dataset introduced by Andersen et al. (1993) and also analyzed by Titman (2015). Patients were randomized to either a treatment of prednisone (251 patients) or a placebo (237 patients). A three-state illness-death model with recovery is assumed, where the healthy and illness states correspond to normal and elevated levels of prothrombin, respectively. A potential measure of the effectiveness of prednisone as a treatment is the restricted conditional expected length of stay in the normal prothrombin level state, from a defined starting point. Specifically, we consider $\Delta \epsilon_{10} = \epsilon_{10}^{(1)}(s, \tau) - \epsilon_{10}^{(2)}(s, \tau)$, the difference in restricted conditional expected length-of-stay in normal prothrombin levels, given the patient is alive and with abnormal prothrombin at time $s$, for the prednisone and placebo groups, corresponding to $\ell = 1$ and $\ell = 2$, respectively. Titman (2015) observed an apparent treatment difference with respect to the transition probabilities with respect to $s = 1000$ days postrandomization. We consider the length of stay in state 0 up to time $\tau = 3000$ days.

In this case, $n_1 = 26$ patients in the prednisone group and $n_2 = 35$ patients in the placebo group meet the condition that $X_i(1000) = 1$ and $Y_i(1000) = 1$. Bootstrap-based confidence intervals are constructed using 1000 bootstrap samples within each group.

Table 3 shows the three constructed 95% confidence intervals, which are broadly similar, although the naive bootstrap interval excludes 0, whereas the Wald and bootstrap-$t$ intervals do not. The AJ-based estimate and a Wald confidence interval is also given. It is seen that there is a dramatic difference between the NM and AJ estimates, with the latter indicating no treatment difference. Potentially, the apparent NM behaviour in the data may be due to patient heterogeneity within the treatment groups.

To illustrate the construction of simultaneous confidence bands, we construct 95% confidence bands for $P_{21}(500, t)$, in the interval $t \in (750, 1250]$. Since it was seen in Section 4 that a reasonable sample size is required for good coverage, an earlier start time is used than for the estimate of the restricted expected length of stay to ensure sufficient numbers of patients are under observation. Since we expect $P_{21}(500, t) < 0.5$, we chose the confidence bands using the transformation $\phi_2(p) = \log(-\log(1 - p))$.

Figure 1 shows the confidence bands for the placebo and prednisone groups using H-W and EP intervals using the NM estimator and using the EP intervals for the AJ estimator.

| Method          | $\Delta \epsilon_{10}$ | 95% CI          |
|-----------------|------------------------|-----------------|
| Wald            | 375.3                  | (-2.1, 752.8)   |
| Naive bootstrap | 375.3                  | (14.7, 740.0)   |
| Bootstrap-$t$   | 375.3                  | (-5.1, 747.3)   |
| AJ (Wald)       | 11.9                   | (-217.3, 241.1) |

Table 3 Estimated difference in conditional length of stay for liver cirrhosis data

Abbreviation: AJ, Aalen-Johansen.
FIGURE 1  Simultaneous confidence bands for the placebo and prednisone groups for the liver cirrhosis dataset using the non-Markov estimator (left panels) and Aalen-Johansen estimator via the wild bootstrap (right panels)

6 | DISCUSSION

To the best of our knowledge, our article is the first to prove weak convergence properties of a general transition probability estimator in independently right-censored NM multistate models while also providing explicit formulas of the asymptotic covariance functions. We gave similar proofs for the classical bootstrap, which finally allows the utilization of the Titman (2015) estimator in time-simultaneous inference procedures. Unfortunately, when focusing on inference on the restricted expected length of stay, both applied bootstrap methods resulted in similarly reliable confidence intervals as the Wald method. In general, sample sizes of at least $n = 100$ ensured a coverage level close to the nominal 95%.

On the other hand, the time-simultaneous confidence bands derived in this article cannot rely on a purely asymptotic quantile finding approach. Instead, they truly require a resampling
method such as the bootstrap method developed at the end of Section 2. Furthermore, the simulation results in Section 4.2 and the application to the liver cirrhosis dataset in Section 5 demonstrate their practical usefulness: not only did they reveal reliable coverage probabilities for relatively small sample sizes already. They were also only slightly wider than the confidence bands based on the AJ estimator for which there is no guarantee of consistency. In future research, one could analyze other resampling procedures as well; they might even improve the performance in small samples for which our NM bands were either too conservative or slightly too liberal. For example, the wild bootstrap methods of Lin, Fleming, and Wei (1994), Lin (1997), Beyersmann, Pauly, and Di Termini (2013), or Bluhmki et al. (2018) are popular choices for resampling counting process-based estimators; see, for example, the frequent use of the multiplier bootstrap with normal variables in the book by Martinussen and Scheike (2006).

In this article, in common with other treatments of NM transition probabilities in the literature, independent censoring has been assumed (Allignol et al., 2014; Meira-Machado et al., 2006; Putter & Spitzni, 2018). For Markov processes, the AJ estimator has the advantage of being robust to censoring which is dependent on the state occupied. Titman (2015) included a simulation study which showed the proposed NM estimator was biased under such scenarios. In the context of estimation of state-occupation probabilities, methods based on inverse probability of censoring weights (IPCWs) have been proposed to allow consistent estimation under state-dependent censoring (Datta & Satten, 2002). In principle, such methods could be applied to the estimator of Titman (2015). However, extending the results of this article to accommodate IPCW-based estimators is not straightforward.

Another field of possible future interest is the extension of the present methodology to quality-adjusted sums of restricted expected lengths of stay in a specific subset of states (cf. Williams, 1985 who proposed such quality-adjusted life years for operation considerations). For instance, the expected length of stay in a healthy state could be weighted with a factor $p \in (0.5, 1)$, whereas the length of stay in the illness state obtains the weight $1 - p$. The specific choice of $p$ may be obtained from additional information on, for example, the pain scores of patients in the illness state and may be chosen individually by the patients themselves. This way, a comparison of treatments could be achieved that more realistically accounts for the circumstances of a specific disease. Estimation of quality-adjusted survival has received substantial attention in the statistical literature. Zhao and Tsiatis (1997, 1999) used the right-censored sample of observed quality-adjusted lifetimes directly to estimate their distribution. Glasziou, Simes, and Gelber (1990) considered a partitioned survival approach to estimate quality-adjusted survival assuming a progressive process. Cook, Lawless, and Lee (2003) compared estimates of quality-adjusted survival based on the AJ and robust Pepe-type estimates of the state occupation probabilities. In addition, quality-adjusted survival estimates based on parametric models making a specific Markov or semi-Markov assumption have been considered by several authors (Chen & Sen, 2001; Pradhan & Dewanji, 2009). Future research, extending the results of this article, could focus on inference on quality- or cost-adjusted life years, possibly simultaneously for multiple different time windows. In this context, resampling methods would again assure the asymptotic exactness of the inference procedure.

However, one should bear in mind the early controversial debate on quality-adjusted life years as articulated by, for example, Harris (1987); see also the discussion of the article by Cox et al. (1992). Therein, G.W. Torrance appreciates the usefulness of quality-adjusted life years for resource allocation if combined with other measures and advices: He states that quality-adjusted life years “were never intended for clinical decision-making—they were developed for use in
resource allocation.” They should also be reported together with other important measures for reaching a decision. Moreover, one should add “a thorough sensitivity analysis and thoughtful discussion, including caveats” to obtain an informative quality-adjusted life years analysis.

Finally, reconsider the data analysis in Section 5, where the AJ estimator yielded a difference of restricted expected lengths of stay, which drastically deviates from the same quantity based on the Titman (2015) estimator. This might be considered as a strong evidence against the Markov assumption. However, as also noted by Putter and Spitoni (2018), if the discrepancy between the AJ and NM estimates is modest, the AJ estimator may be preferred since any bias will likely be outweighed by the substantial gain in efficiency. This is apparent in the mentioned liver cirrhosis data example, where the confidence intervals and simultaneous confidence bands based on the AJ estimator are much narrower. Equivalently, test procedures based on the classical estimator are more powerful in detecting deviations from null hypotheses. As such, there is interest in developing formal tests for the applicability of the classical AJ estimator, which will be the subject of future articles.

ACKNOWLEDGEMENTS
The authors would like to thank the editor, associate editor, and two referees for fruitful discussions which have significantly improved the quality of the article. We also thank Markus Pauly for helpful discussions. Dennis Dobler was supported by a DFG (German Research Foundation) grant.

ORCID
Dennis Dobler https://orcid.org/0000-0002-9040-0854
Andrew Titman https://orcid.org/0000-0003-2084-0720

REFERENCES
Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics, 6*(4), 701–726.
Allignol, A., Beyersmann, J., Gerds, T., & Latouche, A. (2014). A competing risks approach for nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis, 20*(4), 495–513.
Andersen, P. K., Borgan, Ø., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. New York, NY: Springer.
Azarang, L., Scheike, T., & de Uña-Álvarez, J. (2017). Direct modeling of regression effects for transition probabilities in the progressive illness-death model. *Statistics in Medicine, 36*(12), 1964–1976.
Beyersmann, J., Pauly, M., & 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*(3), 387–402.
Bluhmki, T., Schmoor, C., Dobler, D., Pauly, M., Finke, J., Schumacher, M., & Beyersmann, J. (2018). A wild bootstrap approach for the Aalen–Johansen estimator. *Biometrics, 74*(3), 977–985.
Chen, P.-L., & Sen, P. K. (2001). Quality-adjusted survival estimation with periodic observations. *Biometrics, 57*(3), 868–874.
Cook, R. J., Lawless, J. F., & Lee, K.-A. (2003). Cumulative processes related to event histories. *Sort, 27*(1), 0013–0030.
Cox, D. R., Fitzpatrick, R., Fletcher, A. E., Gore, S. M., Spiegelhalter, D. J., & Jones, D. R. (1992). Quality-of-life assessment: can we keep it simple? *Journal of the Royal Statistical Society: Series A, 155*(3), 353–393.
Datta, S., & 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 and Probability Letters, 55*(4), 403–411.
Datta, S., & Satten, G. A. (2002). Estimation of integrated transition hazards and stage occupation probabilities for non-Markov systems under dependent censoring. *Biometrics, 58*(4), 792–802.

Glasziou, P. P., Simes, R. J., & Gelber, R. D. (1990). Quality adjusted survival analysis. *Statistics in Medicine, 9*(11), 1259–1276.

Glidden, D. V. (2002). Robust inference for event probabilities with non-Markov event data. *Biometrics, 58*(2), 361–368.

Gunnes, N., Borgan, Ø., & Aalen, O. O. (2007). Estimating stage occupation probabilities in non-Markov models. *Lifetime Data Analysis, 13*(2), 211–240.

Harris, J. (1987). QALYfying the value of life. *Journal of Medical Ethics, 13*(3), 117–123.

Irwin, J. O. (1949). The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Journal of Hygiene, 47*(2), 188–189.

Grand, M. K., & Putter, H. (2016). Regression models for expected length of stay. *Statistics in Medicine, 35*(7), 1178–1192.

Lin, D. Y. (1997). Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in Medicine, 16*(8), 901–910.

Lin, D. Y., Fleming, T. R., & Wei, L. J. (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika, 81*(1), 73–81.

Martinussen, T., & Scheike, T. H. (2006). *Dynamic regression models for survival data*. New York, NY: Springer.

Meira-Machado, L., de Uña-Álvarez, J., & Cadarso-Suárez, C. (2006). Nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis, 12*(3), 325–344.

Meira-Machado, L., de Uña-Álvarez, J., & Datta, S. (2015). Nonparametric estimation of conditional transition probabilities in a non-Markov illness-death model. *Computational Statistics, 30*(2), 377–397.

Meira-Machado, L., & Sestelo, M. (2019). Estimation in the progressive illness-death model: A nonexhaustive review. *Biometrical Journal, 61*(2), 245–263.

Pepe, M. S. (1991). Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association, 86*(415), 770–778.

Pepe, M. S., Longton, G., & Thornquist, M. (1991). A qualifier Q for the survival function to describe the prevalence of a transient condition. *Statistics in Medicine, 10*(3), 413–421.

Pradhan, B., & Dewanji, A. (2009). Parametric estimation of quality adjusted lifetime (qal) distribution in progressive illness-death model. *Statistics in Medicine, 28*(15), 2012–2027.

Putter, H., & 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.

Rodríguez-Girondo, M., & de Uña-Álvarez, J. (2012). A nonparametric test for markovianity in the illness-death model. *Statistics in Medicine, 31*(30), 4416–4427.

Silber, J. H., Rosenbaum, P. R., Koziol, L. F., Sutaria, N., Marsh, R. R., & Even-Shoshan, O. (1999). Conditional length of stay. *Health Services Research, 34*(1 Pt 2), 349.

Titman, A. C. (2015). Transition probability estimates for non-Markov multi-state models. *Biometrics, 71*(4), 1034–1041.

de Uña-Álvarez, J., & Mandel, M. (2018). Nonparametric estimation of transition probabilities for a general progressive multi-state model under cross-sectional sampling. *Biometrics, 74*(4), 1203–1212.

de Uña-Álvarez, J., & Meira-Machado, L. (2015). Nonparametric estimation of transition probabilities in the non-Markov illness-death model: A comparative study. *Biometrics, 71*(2), 364–375.

Wangler, M., Beyersmann, J., & Schumacher, M. (2006). changeLOS: An R-package for change in length of hospital stay based on the Aalen–Johansen estimator. *The Newsletter of the R Project, 6*(2), 31–35.

Williams, A. (1985). Economics of coronary artery bypass grafting. *British Medical Journal (Clinical Research Edition), 291*(6491), 326–329.

Zhao, H., & Tsiatis, A. A. (1997). A consistent estimator for the distribution of quality adjusted survival time. *Biometrika, 84*(2), 339–348.

Zhao, H., & Tsiatis, A. A. (1999). Efficient estimation of the distribution of quality-adjusted survival time. *Biometrics, 55*(4), 1101–1107.
SUPPORTING INFORMATION
Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Dobler D, Titman A. Dynamic inference for non-Markov transition probabilities under random right censoring. *Scand J Statist*. 2020;47:572–586. https://doi.org/10.1111/sjos.12443