Published in final edited form as:
Can J Stat. 2013 June ; 41(2): . doi:10.1002/cjs.11176.

Estimation with Right-Censored Observations Under A Semi-Markov Model

Lihui Zhao[1] and **X. Joan Hu**[2]

Lihui Zhao: lihui.zhao@northwestern.edu

[1]Department of Preventive Medicine, Northwestern University, Chicago, IL 60611, USA

[2]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC V5A1S6, Canada

Abstract

The semi-Markov process often provides a better framework than the classical Markov process for the analysis of events with multiple states. The purpose of this paper is twofold. First, we show that in the presence of right censoring, when the right end-point of the support of the censoring time is strictly less than the right end-point of the support of the semi-Markov kernel, the transition probability of the semi-Markov process is nonidentifiable, and the estimators proposed in the literature are inconsistent in general. We derive the set of all attainable values for the transition probability based on the censored data, and we propose a nonparametric inference procedure for the transition probability using this set. Second, the conventional approach to constructing confidence bands is not applicable for the semi-Markov kernel and the sojourn time distribution. We propose new perturbation resampling methods to construct these confidence bands. Different weights and transformations are explored in the construction. We use simulation to examine our proposals and illustrate them with hospitalization data from a recent cancer survivor study.

Keywords

Case fatality ratio; Confidence band; Identifiability; Multi-state process; Semi-Markov kernel; Semi-Markov process; Sojourn time distribution; Transition probability

1 Introduction

Multi-state stochastic processes provide a convenient framework for the analysis of event history data (Andersen et al., 1993; Commenges, 1999; Andersen and Keiding, 2002). The semi-Markov model is a generalization of the classical homogeneous Markov model. It is preferred in many practical applications because it accommodates the dependence of transitions between states on the state durations (Andersen, Esbjerg, and Sorensen, 2000; Kang and Lagakos, 2007).

Suppose $\{S(t):t \geq 0\}$ is a finite-state stochastic process with state space $\mathscr{E} = \{1,\ldots, r\}$. Denote its sequence of consecutive states by $\{J_m:m = 0, 1, 2,\ldots\}$, and let the corresponding transition times be $\{T_m:m = 0, 1,\ldots\}$. Without loss of generality, we assume $T_0 = 0$. Let $X_m = T_m - T_{m-1}$ be the sojourn time of the mth transition. Then $\{S(t):t \geq 0\}$ is a homogeneous

Correspondence to: Lihui Zhao, lihui.zhao@northwestern.edu.

semi-Markov process if for any state $j \in \mathscr{E}$ and duration time $\tau > 0$, the transition probability satisfies

$$P\{J_{m+1}=j, X_{m+1} \leq \tau | J_0, T_0, \ldots, J_m, T_m\} = P\{J_{m+1}=j, X_{m+1} \leq \tau | J_m\}, \quad (1)$$

and (1) is independent of m (Ross, 1996). Denote the function in (1) with $J_m = h (\neq j)$ by $Q_{hj}(\tau)$. The set $\{Q_{hj}(\cdot): h, j \in \mathscr{E}\}$, referred to as the semi-Markov kernel, fully characterizes the corresponding homogeneous semi-Markov process given its first state $S(0) = J_0$. Lagakos, Sommer, and Zelen (1978) present the nonparametric maximum likelihood estimator (MLE) of the semi-Markov kernel $Q_{hj}(\cdot)$ based on right-censored observations. The consistency and weak convergence of their nonparametric MLE are derived by Gill (1980) using the theory of stochastic integration and counting processes. Matthews (1984) and Dinse and Larson (1986) express the nonparametric MLE in terms of cause-specific hazard functions, which simplify its calculation and clarify its interpretation.

In a homogeneous semi-Markov process $\{S(t), t \geq 0\}$, the sequence of its consecutive states $\{J_m = S(T_m): m = 0, 1, 2, \ldots\}$ is an embedded homogeneous Markov chain with the transition probability

$$P_{hj} = P\{J_{m+1}=j | J_m = h\} = \lim_{\tau \to \infty} Q_{hj}(\tau), h, j \in \mathscr{E}. \quad (2)$$

If h is an absorbing state, we have $Q_{hj}(\cdot) = 0$ and therefore $P_{hj} = 0$ for every $j \neq h$. When h is not absorbing, the sojourn time between h and j has the cumulative distribution function

$$F_{hj}(\tau) = P\{X_{m+1} \leq \tau | J_m = h, J_{m+1} = j\} = Q_{hj}(\tau)/P_{hj}, \tau > 0. \quad (3)$$

It is often of interest to estimate the semi-Markov transition probability P_{hj} defined in (2) and the sojourn time distribution $F_{hj}(\tau)$ in (3).

Lagakos et al. (1978) propose two estimators for P_{hj}: a plug-in estimator based on the nonparametric MLE of the semi-Markov kernel, and its normalized version. However, the large sample properties of their estimators have not been carefully studied. In Section 2, we show that the convergent limits (in probability) of the two estimators are different from P_{hj} when the right end-point of the support of the censoring time is strictly less than the right end-point of the support of $Q_{hj}(\cdot)$. Phelan (1990) proposes an estimator for P_{hj} using the proportion of completely observed "$h \to j$" transitions among all the uncensored transitions starting from state h. His estimator discards all the incompletely observed transitions due to censoring. He establishes the consistency and asymptotic normality of his estimator under the assumption that all the sojourn times in each of the states have the same distribution regardless of the next state that the process transits to, i.e.,

$$F_{hj}(\cdot) = H_h(\cdot) \text{for} j \in \mathscr{E} \quad (4)$$

where

$$H_h(\tau) = P\{X_{m+1} \leq \tau | J_m = h\} = \sum_{k \neq h} Q_{hk}(\tau), \quad (5)$$

is the distribution of the sojourn time in state h regardless of the next state. However, when assumption (4) is violated, the sojourn time distributions $F_{hj}(\cdot)$ depend on the next states, and transitions associated with longer sojourn times are more likely to be censored. Discarding the information contained in censored transitions may thus result in a biased

estimator for the transition probabilities P_{hj}. This is confirmed by the simulation study reported in Section 4.

However, in practice, the follow-up time is usually finite and the right end-point of its support can be strictly less than the right end-point of the support of $Q_{hj}(\cdot)$. In addition, assumption (4) can be inappropriate because the sojourn time distributions $F_{hj}(\cdot)$ likely depend on j. This can be exemplified by the study of childhood cancer survivors reported inMcBride et al. (2010). The CAYACS (Childhood, Adolescent, and Young Adult Survivors) study database includes hospitalization records and death information for its subjects from 1986 to 2000. If we model the hospitalization process by a semi-Markov model with the states “health”, “in hospital”, and “death”, it is unlikely that the sojourn times for transitions from “health” to “in hospital” have the same distribution as those for transitions from “health” to “death”. In fact, a preliminary analysis of these data under assumption (4) yielded some rather counterintuitive results; see Section 5 for details. Thus, for this example, the existing estimators for transition probabilities can be inconsistent and misleading. This partly motivates this research. In Section 2, we develop a new estimation procedure for the transition probabilities P_{hj} that is valid even when assumption (4) is violated.

Moreover, as pointed out by Gill (1980), the asymptotic Gaussian process of the non-parametric MLE of the semi-Markov kernel $Q_{hj}(\cdot)$ does not have an independent increment structure. Thus, the limiting process can not be transformed by smooth maps into the standard Brownian bridge or Brownian motion. The well-established procedures for constructing confidence bands, such as that presented in Hall and Wellner (1980), are thus not directly applicable to the semi-Markov kernel. In Section 3, we develop a new perturbation-resampling method to overcome this difficulty. Similar techniques have been used successfully in other contexts, for example by Lin, Wei, and Ying (1993). Similarly to the transition probability P_{hj}, the sojourn time distribution $F_{hj}(\cdot)$ is also nonidentifiable in general. Using the relationship that $F_{hj}(\tau) = Q_{hj}(\tau) / P_{hj}$, in Section 3 we propose a method to construct confidence bands for $F_{hj}(\cdot)$.

Thus, the purpose of this paper is twofold. We first propose a new inference procedure for the transition probability P_{hj} under the semi-Markov model with right-censored observations. Secondly, we present resampling-based methods to construct confidence bands for the semi-Markov kernel function $Q_{hj}(\cdot)$ and the sojourn time distribution $F_{hj}(\cdot)$. We organize the rest of the paper as follows. Section 2 introduces the framework, derives the convergent limits (in probability) of the existing estimators for the transition probability P_{hj}, and shows the inconsistency of the existing estimators when assumption (4) does not hold. We then develop a new estimation procedure that is valid even when assumption (4) is violated. Section 3 presents resampling-based methods to construct confidence bands for the semi-Markov kernel function and the associated sojourn time distribution. We explore different weights and transformations in the construction. Section 4 reports the results of simulation studies conducted to examine the finite sample performance of the proposed approaches. Section 5 provides an analysis of the hospitalization data using the proposed approaches along with the preliminary analysis mentioned above. Section 6 provides some concluding remarks.

2 Estimation of Transition Probability P_{hj}

We begin with a description of the framework and then briefly review the existing estimators for the transition probability defined in (2) with right-censored semi-Markov process data. We show that with right-censored observations the estimators can be

inconsistent and the transition probability can be nonidentifiable. We then propose an alternative estimation procedure.

2.1 Framework

Recall that $\{J_m: m = 0, 1, 2,\ldots\}$ is the sequence of consecutive states of the semi-Markov process $S(\cdot)$, X_m is the sojourn time of the mth transition, and T_m is the time of the mth transition. We follow the counting-process formulation presented in Gill (1980). Given a study time period $[0, t]$, for any $u \geq 0$, let

$$N_{hj}(u;t)=\#\{m:m \geq 1, J_{m-1}=h, J_m=j, X_m \leq u, T_m \leq t\}$$

be the number of sojourn times in state h that are $\leq u$ and followed by a transition to state j, and let

$$Y_h(u;t)=\#\{m:m \geq 1, J_{m-1}=h, X_m \geq u, T_{m-1}+u \leq t\}$$

be the number of sojourn times in state h that are $\geq u$. Note that both $N_{hj}(u;t)$ and $Y_h(u;t)$ are 0 if $u >$ t.

In practice, the observation of a semi-Markov process $S(\cdot)$ is usually subject to right censoring. We assume throughout the paper that the censoring time C is independent of $S(\cdot)$. Suppose we have n independent realizations from right-censored observations of $S(\cdot)$, denoted by $\{S_i(t), t \in [0, C_i]\}_{i=1}^n$. For the ith realization, we observe $\{N_{hji}(u;C_i), u \geq 0\}$ and $\{Y_{hi}(u;C_i), u \geq 0\}$, which will be denoted by $\{N_{hji}(u), u \geq 0\}$ and $\{Y_{hi}(u), u \geq 0\}$ for simplicity of notation. Following the notation of Gill (1980), we let $N_{hj}^n(\cdot)=\sum_{i=1}^n N_{hji}(\cdot)$, $N_h^n(\cdot)=\sum_{j\in\mathcal{E}} N_{hj}^n(\cdot)$, and $Y_h^n(\cdot)=\sum_{i=1}^n Y_{hi}(\cdot)$. The nonparametric MLE derived in Lagakos et al. (1978) for the semi-Markov kernel $Q_{hj}(\cdot)$ defined in (1) can be written

$$\widehat{Q}_{hj}(\tau)=\int_0^\tau \{1-\widehat{H}_h(u-)\}\frac{dN_{hj}^n(u)}{Y_h^n(u)}, \tau \geq 0, \quad (6)$$

where

$$\widehat{H}_h(u)=1-\prod_{v\leq u}\left\{1-\frac{dN_h^n(v)}{Y_h^n(v)}\right\}, u \geq 0$$

is the nonparametric MLE of $H_h(\cdot)$ defined in (5). Gill (1980) establishes the uniform consistency and weak convergence of $\{\hat{Q}_{hj}(\tau); h, j \in \mathcal{E}\}$ over a range of τ. We consider the estimation of the semi-Markov transition probability P_{hj} defined in (2).

2.2 Study of existing estimators

Because $P_{hj} = \lim_{\tau\to\infty} Q_{hj}(\tau)$, Lagakos et al. (1978) suggest a plug-in estimator of the transition probability: $\hat{P}_{hj} = \hat{Q}_{hj}(\infty)$ for $h, j \in \mathcal{E}$. We refer to this as the LSZ plug-in estimator.

The LSZ plug-in estimator $\hat{Q}_{hj}(\infty)$ is in fact the same as the MLE $\hat{Q}_{hj}(\cdot)$ evaluated at the largest observed sojourn time starting from state h. It is thus possible that $\sum_j \hat{P}_{hj} < 1$ with

right-censored observations, which is undesirable in practical applications. Hence,Lagakos et al. (1978) propose a normalized version of the plug-in estimator: $\tilde{P_{hj}} = \hat{Q_{hj}}(\infty)/\Sigma_{k\neq h} \hat{Q_{hk}}(\infty)$. We refer to this as the LSZ normalized estimator.

Denote the largest sojourn time in state h, $h \in \mathscr{E}$, that is potentially observable by

$$\tau_h = \sup\{\tau : P(Y_h(\tau) > 0) > 0\}. \quad (7)$$

The following proposition presents the limiting properties of the LSZ estimators.

Proposition 1—*Assume that the semi-Markov kernel function $Q_{hj}(\cdot)$ in (1) is continuous for all $h, j \in \mathscr{E}$. Also assume $E\{Y_h(0)\} < \infty$ for all $h \in \mathscr{E}$. Then $P(Q_{hj}(\tau_h) = \hat{Q_{hj}}(\infty)) = 1$ for all $h, j \in \mathscr{E}$. Moreover, as $n \to \infty$, the LSZ plug-in estimator $\hat{P_{hj}}$ and the LSZ normalized estimator $\tilde{P_{hj}}$ converge in probability to $Q_{hj}(\tau_h)$ and $Q_{hj}(\tau_h)/\Sigma_{k\neq h} Q_{hk}(\tau_h)$, respectively.*

The proof is outlined in Appendix A.1. The condition $E\{Y_h(0)\} < \infty$ indicates that the number of at least partially observed sojourn times is almost surely finite. This proposition reveals the following:

- **i.** The LSZ plug-in estimator $\hat{P_{hj}}$ is a consistent estimator for the transition probability P_{hj} if and only if $Q_{hj}(\tau_h) = Q_{hj}(\infty) = P_{hj}$.
- **ii.** The LSZ normalized estimator $\tilde{P_{hj}}$ is a consistent estimator for P_{hj} if and only if $P_{hj} = Q_{hj}(\tau_h)/\Sigma_{k\neq h} Q_{hk}(\tau_h)$.

In general, both $\hat{P_{hj}}$ and $\tilde{P_{hj}}$ are inconsistent. For example, when the right end-point of the support of the censoring time is strictly less than the right end-point of the support of $Q_{hj}(\cdot)$, it is straightforward to show that $Q_{hj}(\tau_h) < Q_{hj}(\infty) = P_{hj}$. It follows that the LSZ plug-in estimator $\hat{P_{hj}}$ is not consistent for P_{hj}. For the LSZ normalized estimator $\tilde{P_{hj}}$, it is easy to verify that the condition in (ii) above holds, provided $F_{hj}(\tau_h) = H_h(\tau_h)$ for $j \in \mathscr{E}$. On the other hand, if at least one $F_{hj}(\tau_h)$, $j \in \mathscr{E}$ is different from the others for a fixed h, then at least one $F_{hk}(\tau_h)/F_{hj*}(\tau_h)$, $k \in \mathscr{E}$ is larger than 1, where $F_{hj*}(\tau_h) = \min_{j\in\mathscr{E}} F_{hj}(\tau_h)$. It follows that

$$\sum_{k\neq h} Q_{hk}(\tau_h) = F_{hj*}(\tau_h) \sum_{k\neq h} P_{hk} \frac{F_{hk}(\tau_h)}{F_{hj*}(\tau_h)} > F_{hj*}(\tau_h).$$

Thus, $\tilde{P_{hj*}}$ converges in probability to $Q_{hj*}(\tau_h)/\Sigma_{k\neq h} Q_{hk}(\tau_h) < P_{hj*}$. That is, the condition (ii) required for the consistency of the LSZ normalized estimator $\tilde{P_{hj}}$ is equivalent to $F_{hj}(\tau_h) = H_h(\tau_h)$ for $j \in \mathscr{E}$, which is necessary for assumption (4). However, this condition is often violated, and thus $\tilde{P_{hj}}$ is not consistent for P_{hj}. In the next subsection, we propose a nonparametric inference procedure for P_{hj} that does not rely on these conditions.

2.3 Proposed interval estimator

It follows from the definition of τ_h that the semi-Markov kernel $Q_{hj}(\cdot)$ is estimable only up to τ_h. The available information, the right-censored observations of the semi-Markov process, is on $Q_{hj}(\tau)$ for $\tau \in [0, \tau_h]$ with $h, j \in \mathscr{E}$. Although the transition probability P_{hj} is not identifiable in general, the set of its attainable values is well defined. Let $P_{hj}^L = Q_{hj}(\tau_h)$ and $P_{hj}^U = 1 - \sum_{j'\neq j} Q_{hj'}(\tau_h)$. Note that $Q_{hj}(\tau_h) = P_{hj}F_{hj}(\tau_h) \leq P_{hj}$ and $\Sigma_{j'\neq j} Q_{hj'}(\tau_h) \leq \Sigma_{j'\neq j} P_{hj'} = 1 - P_{hj}$. The transition probability P_{hj} is thus contained in $[P_{hj}^L, P_{hj}^U]$, the set of all attainable values of P_{hj}. That is, the data provide information about P_{hj} only through the information on the two interval limits, P_{hj}^L and P_{hj}^U. When the right end-point of the support

of the censoring time is strictly less than the right end-point of the support of $Q_{hj}(\cdot)$, it is straightforward to show that $P_{hj}^L < P_{hj}^U$. Without further assumptions, P_{hj} is therefore not identifiable based on the data. By Proposition 1, $\widehat{P}_{hj}^L = \widehat{Q}_{hj}(\infty)$ and $\widehat{P}_{hj}^U = 1 - \sum_{j' \neq j} \widehat{Q}_{hj'}(\infty)$ are consistent estimators of P_{hj}^L and P_{hj}^U, respectively. This motivates the following procedure for constructing a confidence interval for P_{hj}.

Note that a $(1 - \alpha)$ confidence interval for any value in the interval $[P_{hj}^L, P_{hj}^U]$ can be constructed by determining two positive constants c_1 and c_2 such that

$$P\left([P_{hj}^L, P_{hj}^U] \subseteq [\widehat{P}_{hj}^L - c_1, \widehat{P}_{hj}^U + c_2]\right) \geq 1 - \alpha. \quad (8)$$

Since $P_{hj} \in [P_{hj}^L, P_{hj}^U]$, we propose estimating the transition probability P_{hj} with $[\widehat{P}_{hj}^L - c_1, \widehat{P}_{hj}^U + c_2]$, the interval from (8). This yields a confidence interval of P_{hj} with level at least $1 - \alpha$, without any assumption beyond those of Proposition 1.

Based on the asymptotic properties of the MLE $\hat{Q}_{hj}(\cdot)$ derived in Gill (1980), the distribution of $\{\hat{Q}_{hj}(\tau_h): h, j \in \mathscr{E}\}$ is asymptotically normal with mean $\{Q_{hj}(\tau_h): h, j \in \mathscr{E}\}$ provided $P(Y_h(\tau_h) > 0) > 0$. The joint distribution of $(\widehat{P}_{hj}^L, \widehat{P}_{hj}^U)$ with $h, j \in \mathscr{E}$, a bivariate function of $(\hat{Q}_{hl}(\tau_h): l \in \mathscr{E})$ with continuous derivatives, is thus asymptotically normal with mean (P_{hj}^L, P_{hj}^U). Theoretically speaking, c_1 and c_2 in (8) can be determined by the asymptotic distribution of $(\widehat{P}_{hj}^L, \widehat{P}_{hj}^U)'$. However, the limiting distribution is not convenient to use analytically for this purpose. Instead, we approximate the limiting distribution using an innovative nonparametric bootstrap procedure. This approximation also allows us to determine many pairs of c_1 and c_2 that satisfy (8). We suggest using the pair that minimizes $c_1 + c_2$ because this leads to the shortest confidence interval. This optimization can be easily done using the bootstrap approximation. This will be illustrated and discussed in Section 4 with a simulation study and in Section 5 with the aforementioned hospitalization data.

The proposed confidence interval for P_{hj} results from an interval estimator for $[P_{hj}^L, P_{hj}^U]$, the set of attainable values of P_{hj}. It is usually wider than the corresponding confidence intervals based on existing estimators. Note that the LSZ plug-in estimator $\hat{P}_{hj}$ is the same estimator $\widehat{P}_{hj}^L$ for the lower bound P_{hj}^L. The LSZ normalized estimator $\tilde{P}_{hj}$ takes values between $\widehat{P}_{hj}^L$ and $\widehat{P}_{hj}^U$. However, our confidence interval does not require assumption (4), while confidence intervals based on existing estimators can have a level much lower than the nominal level because of the possible inconsistency of the estimators when the assumption (4) is violated. On the other hand, when the data contain more information about the attainable values of P_{hj}, $[P_{hj}^L, P_{hj}^U]$ becomes narrower, and our confidence interval is comparable to those based on existing estimators. We will further discuss the robustness and efficiency of the proposed interval estimator with the numerical results in Sections 4 and 5.

3 Estimation of Semi-Markov Kernel and Sojourn Time Distribution

This section focuses on procedures for constructing confidence bands for the semi-Markov kernel and the sojourn time distribution.

3.1 Confidence band for semi-Markov kernel

Gill (1980) points out that the limiting process of $n^{1/2}\{\hat{Q}_{hj}(\cdot) - Q_{hj}(\cdot)\}$ does not have the independent-increment structure, and thus it can not be transformed into the standard Brownian bridge or Brownian motion. Therefore, the conventional approach to constructing confidence bands for an unknown function is not applicable. We propose a new perturbation-resampling method to construct confidence bands for the semi-Markov kernel.

With $h, j \in \mathscr{E}$ and $\tau > 0$, define

$$\hat{Z}_{hji}(u) = N_{hji}(u) - \int_0^u \frac{Y_{hi}(s)}{Y_h^n(s)} dN_{hj}^n(s), u > 0$$

and

$$W_{hj}^n(\tau) = n^{1/2} \sum_{i=1}^n U_i \left\{ \int_0^\tau \frac{1 - \hat{H}_h(s-)}{Y_h^n(s)} d\hat{Z}_{hji}(s) + \int_0^{\tau-} \frac{(\hat{Q}_{hj}(s) - \hat{Q}_{hj}(\tau))(1 - \hat{H}_h(s-))}{(1 - \hat{H}_h(s)) Y_h^n(s)} d \sum_{k \in \mathscr{E}} \hat{Z}_{hki}(s) \right\}, \quad (9)$$

where U_i, $i = 1, \ldots, n$ are independent standard normal random variables.

Proposition 2—*Assume that the semi-Markov kernel $Q_{hj}(\cdot)$ is continuous for $h, j \in \mathscr{E}$ and $E\{Y_h(0)^{7+\varepsilon}\} < \infty$ for some $\varepsilon > 0$. Let ν_h be a constant satisfying $P(Y_h(\nu_h) > 0) > 0$. Then, conditional on the available data, $\{W_{hj}^n(\tau): \tau \in [0, \nu_h]\}$ and $\{n^{1/2}\{\hat{Q}_{hj}(\tau) - Q_{hj}(\tau)\}: \tau \in [0, \nu_h]\}$ converge weakly to the same Gaussian process as $n \to \infty$ for all $h, j \in \mathscr{E}$.*

The proof is outlined in Appendix A.2.

Various confidence bands for $Q_{hj}(\cdot)$ can be constructed using the class of transformed Processes

$$G_{hj}(\tau) = n^{1/2} g_{hj}^{(n)}(\tau)[\phi(\hat{Q}_{hj}(\tau)) - \phi(Q_{hj}(\tau))],$$

where $\varphi(\cdot)$ is a fixed function with its first derivative $\varphi'(\cdot)$ nonzero and continuous, and $g_{hj}^{(n)}(\cdot)$ is a weight function based on the available data with a deterministic limit $g_{hj}(\tau)$ in probability. The weight $g_{hj}^{(n)}(\cdot)$ determines the shape of the bands. By the functional delta-method (e.g., Andersen et al., 1993), the process $G_{hj}(\cdot)$ is asymptotically equivalent to

$$g_{hj}^{(n)}(\tau)\phi'(\hat{Q}_{hj}(\tau))\{n^{1/2}[\hat{Q}_{hj}(\tau) - Q_{hj}(\tau)]\}, \tau > 0.$$

In practice, one often uses a transformation with positive $\varphi'(\cdot)$. A natural transformation function is the identity function $\varphi(x) = x$. We also consider another transformation function, $\varphi(x) = \log(-\log(1 - x))$. This is analogous to the function widely used in the literature to improve the coverage probabilities of confidence intervals and confidence bands for a survival function with the Kaplan–Meier estimator. Another advantage of the transformation is that it ensures that the bounds are between 0 and 1, as required for a survival function. We consider two choices for the weight function for the identity transformation function $g_{hj}^{(n)}(\tau)$:

$$\{1-\widehat{Q}_{hj}(\tau)\}\log\{1-\widehat{Q}_{hj}(\tau)\}/\widehat{\sigma}_{hj}(\tau) \quad (10)$$

for the identity transformation function $\varphi(x) = x$, and

$$\log\{1-\widehat{Q}_{hj}(\tau)\}/\left\{1+\widehat{\sigma}^2_{hj}(\tau)/\{1-\widehat{Q}_{hj}(\tau)\}^2\right\} \quad (11)$$

for the transformation function $\varphi(x) = \log(-\log(1 - x))$, where $\widehat{\sigma}^2_{hj}(\tau)$ is a consistent estimator for the asymptotic variance of $n^{1/2}\{\widehat{Q}_{hj}(\tau) - Q_{hj}(\tau)\}$. When $P_{hj} = 1$, $Q_{hj}(\cdot)$ is a distribution function, and the resulting confidence bands for $Q_{hj}(\cdot)$ reduce to the widely used equal-precision (EP) bands (Nair, 1984) for weight (10) and the Hall–Wellner (HW) bands (Hall and Wellner, 1980) for weight (11), respectively. We therefore refer to the confidence bands for weights (10) and (11) as EP and HW bands.

By Proposition 2, we can approximate the critical values required in the construction of a $(1-\alpha)$ confidence band for $Q_{hj}(\cdot)$ over $[s_1, s_2]$, a predetermined subset of $[0, \nu_h]$. For a given transformation $\varphi(\cdot)$, the limits of an approximate $(1 - \alpha)$ confidence band for $\varphi(Q_{hj}(\tau))$ on $[s_1, s_2]$ are

$$\phi(\widehat{Q}_{hj}(\tau)) \pm n^{-1/2} q_{hj}(s_1, s_2)/g^{(n)}_{hj}(\tau), \tau \in [s_1, s_2], \quad (12)$$

where $q_{hj}(s_1, s_2)$ is the $(1 - \alpha)$ quantile of $q^{(b)}_{hj}(s_1, s_2)$:$b=1, \ldots B\}$; these values are obtained as follows:

Step 1. Generate B sets of independent standard normal random variables $U^{(b)}_i$:$i=1 \ldots n\}$ for $b = 1,\ldots, B$, and obtain the corresponding realizations of $W^n_{hj}(\cdot)$, denoted by $\{W^{n(b)}_{hj}(\cdot)$:$b=1, \ldots, B\}$.

Step 2 For $b = 1,\ldots,B$, let $G^{(b)}_{hj}(\tau)=g^{(n)}_{hj}(\tau)\phi'(\widehat{Q}_{hj}(\tau))W^{n(b)}_{hj}(\tau)$ and obtain $q^{(b)}_{hj}(s_1, s_2)=\sup_{\tau\in[s_1,s_2]}|G^{(b)}_{hj}(\tau)|$.

The limits of a confidence band for $Q_{hj}(\cdot)$ on $[s_1, s_2]$ can then be obtained from (12) by converting the transformation $\varphi(\cdot)$.

3.2 Confidence band for sojourn time distribution

Noting that $F_{hj}(\tau) = Q_{hj}(\tau)/P_{hj}$, similarly to the transition probability P_{hj}, the sojourn time distribution $F_{hj}(\cdot)$ is also nonidentifiable in general. Let $F^L_{hj}(\cdot)=Q_{hj}(\cdot)/P^U_{hj}$ and $F^U_{hj}(\cdot)=Q_{hj}(\cdot)/P^L_{hj}$. Since the sojourn time distribution $F_{hj}(\cdot)$ is bounded by the two functions $F^L_{hj}(\cdot)$ and $F^U_{hj}(\cdot)$, we may obtain a confidence band for $F_{hj}(\cdot)$ on a given interval, say, $[s_1, s_2]$, by constructing a confidence band for $[F^L_{hj}(\tau), F^U_{hj}(\tau)]$, $\tau \in [s_1, s_2]$. To ensure that the confidence band lies within [0, 1] as desired, we consider a transformation $\varphi(\cdot)$, such as $\varphi(x) = \log(-\log(1 - x))$ as in Section 3.1. This may also improve the coverage of the confidence bands. Define

$$D^L_{hj}(\tau)=n^{1/2}d^{1n}_{hj}(\tau)\phi'(\widehat{F}^L_{hj}(\tau))\{\widehat{F}^L_{hj}(\tau) - F^L_{hj}(\tau)\}, \tau>0 \quad (13)$$

and

$$D_{hj}^{U}(\tau)=n^{1/2}d_{hj}^{2n}(\tau)\phi'(\widehat{F}_{hj}^{U}(\tau))\{\widehat{F}_{hj}^{U}(\tau)-F_{hj}^{U}(\tau)\},\tau>0, \quad (14)$$

where $d_{hj}^{1n}(\tau)$ and $d_{hj}^{2n}(\tau)$ are weight functions with deterministic limits $d_{hj}^{1}(\tau)$ and $d_{hj}^{2}(\tau)$ in probability, respectively, and $\widehat{F}_{hj}^{L}(\tau)=\widehat{Q}_{hj}(\tau)/\widehat{P}_{hj}^{U}$ and $\widehat{F}_{hj}^{U}(\tau)=\widehat{Q}_{hj}(\tau)/\widehat{P}_{hj}^{L}$. We may use the weight functions analogous to (10) and (11), which produce EP and HW confidence bands. Choose the quantities $q_{hj}^{L}(s_1,s_2)$ and $q_{hj}^{U}(s_1,s_2)$ in

$$\left[\phi(\widehat{F}_{hj}^{L}(\tau))-n^{-1/2}q_{hj}^{L}(s_1,s_2)/d_{hj}^{1n}(\tau),\phi(\widehat{F}_{hj}^{U}(\tau))+n^{-1/2}q_{hj}^{U}(s_1,s_2)/d_{hj}^{2n}(\tau)\right],\tau\in[s_1,s_2] \quad (15)$$

such that (15) contains $[\phi(F_{hj}^{L}(\tau)),\phi(F_{hj}^{U}(\tau))]$, $\tau\in[s_1,s_2]$ with probability at least $1-\alpha$. This gives a $(1-\alpha)$ confidence band for $F_{hj}(\cdot)$ over the interval $[s_1,s_2]$ by converting the transformation $\varphi(\cdot)$ for the two bounds in the confidence band (15)

The efficiency of the proposed confidence band for $F_{hj}(\cdot)$ over a predetermined interval is usually not very high, since it is converted from a confidence band for $[\phi(F_{hj}^{L}(\cdot)),\phi(F_{hj}^{U}(\cdot))]$ with the same level. We may provide a confidence band with improved efficiency when additional information is available or a further assumption is made.

Recalling the definition of the estimators $\widehat{P}_{hj}^{L}$ and $\widehat{P}_{hj}^{U}$ given in Section 2.3, $\widehat{F}_{hj}^{L}(\tau),\widehat{F}_{hj}^{U}(\tau))$ is a bivariate function of $\hat{Q}_{hj}(\tau)$ and $\{\hat{Q}_{hl}(\tau_h):l\in\mathcal{E}\}$ with continuous derivatives. By the weak convergence of $\hat{Q}_{hj}(\cdot)$ and the fact that $\hat{Q}_{hl}(\tau_h)$ converges in probability to $P_{hl}^{L}=Q_{hl}(\tau_h)$ (see Proposition 1), $n^{1/2}(\widehat{F}_{hj}^{L}(\cdot)-F_{hj}^{L}(\cdot),\widehat{F}_{hj}^{U}(\cdot)-F_{hj}^{U}(\cdot))'$ converges weakly to a bivariate mean zero Gaussian process. So does $(D_{hj}^{L}(\cdot),D_{hj}^{U}(\cdot))'$ for an appropriate transformation $\varphi(\cdot)$. Theoretically speaking, the critical values $q_{hj}^{L}(s_1,s_2)$ and $q_{hj}^{U}(s_1,s_2)$ in (15) can be determined based on the joint limiting distribution of $(D_{hj}^{L}(\tau),D_{hj}^{U}(\tau))$, $\tau\in[s_1,s_2]$. This is in general hard to implement analytically. We propose the following algorithm for approximating $q_{hj}^{L}(s_1,s_2)$ and $q_{hj}^{U}(s_1,s_2)$ using a nonparametric bootstrap approach.

Step 1. Randomly select a sample $\mathcal{M}$ of size n with replacement from $\{(N_{hji}(\cdot),Y_{hi}(\cdot)):i=1,\ldots,n\}$, and evaluate $\hat{Q}_{hj}(\cdot)$, $\widehat{P}_{hj}^{L}$, $\widehat{P}_{hj}^{U}$, and thus $\widehat{F}_{hj}^{L}(\tau)$ and $\widehat{F}_{hj}^{U}(\tau)$ based on the resampled data $\mathcal{M}$. Repeat this procedure B times to obtain $\{\widehat{F}_{hj}^{(b)L}(\cdot),\widehat{F}_{hj}^{(b)U}(\cdot):b=1,\ldots,B\}$.

Step 2. For $b=1,\ldots,B$, substitute $\widehat{F}_{hj}^{(b)L}(\cdot)$ and $\widehat{F}_{hj}^{(b)U}(\cdot)$ into (13) and (14) to obtain $D_{hj}^{(b)L}(\tau)$ and $D_{hj}^{(b)U}(\tau)$. Let $q_{hj}^{(b)L}(s_1,s_2)=\sup_{\tau\in[s_1,s_2]}D_{hj}^{(b)L}(\tau)$ and $q_{hj}^{(b)U}(s_1,s_2)=\inf_{\tau\in[s_1,s_2]}D_{hj}^{(b)U}(\tau)$. Determine $q_{hj}^{L}(s_1,s_2)$ and $q_{hj}^{U}(s_1,s_2)$ as the $100\alpha_1\%$ quantile of $\{q_{hj}^{(b)L}(s_1,s_2):b=1,\ldots,B\}$ and the $100(1-\alpha_2)\%$ quantile of $\{q_{hj}^{(b)U}(s_1,s_2):b=1,\ldots,B\}$, respectively, with $\alpha_1+\alpha_2=\alpha$.

We may choose α_1 and α_2 in Step 2 to optimize the width of the resulting confidence band.

4 Simulation Study

To examine the finite sample behavior of the proposed estimators, we compared them via simulation with the estimators given in Lagakos et al. (1978) and Phelan (1990).

4.1 The Setting

We conducted a simulation with a three-state semi-Markov process, which is equally likely to start from state 1 or state 2 and has state 3 as an absorbing state. The transition probabilities of the embedded Markov chain were set to $P_{12} = 0.7$, $P_{13} = 0.3$, and $P_{21} = P_{23} = 0.5$, and the sojourn time distributions $F_{hj}(\cdot)$ were set to either:

Setting 1. $F_{12} = exp(2)$, $F_{13} = exp(2)$, $F_{21} = exp(1)$, $F_{23} = exp(1)$.

Setting 2. $F_{12} = exp(1)$, $F_{13} = exp(2)$, $F_{21} = exp(1)$, $F_{23} = exp(1)$.

Setting 3. $F_{12} = exp(1)$, $F_{13} = U(0, 2)$, $F_{21} = U(0, 2)$, $F_{23} = exp(2)$.

Here $exp(a)$ and $U(c, d)$ represent the exponential distribution with mean a and the uniform distribution with parameters c and d, respectively. In each of the three settings, a total of n independent realizations of the semi-Markov processes were simulated and observed subject to noninformative right censoring. The censoring times were generated independently from the uniform distribution $U(0, c_{max})$ with $c_{max} = 3$ or 5, and the sample size n was 50, 100, or 200.

The simulation settings were chosen to study the performance of our estimators in comparison with the existing estimators in various situations. In Setting 1, because $F_{12} = F_{13}$ and $F_{21} = F_{23}$, the LSZ normalized and Phelan estimators are consistent. However, the LSZ plug-in estimator will be biased because $Q_{hj}(\tau_h) < Q_{hj}(\infty)$. The bias will be bigger when the maximum censoring time $c_{max} = 3$, which is relatively small. Setting 2 has $F_{21} = F_{23}$ but $F_{12} \neq F_{13}$. Thus, the LSZ normalized and Phelan estimators will perform well for P_{21} and P_{23} but not necessarily for P_{12} and P_{13}. Similarly to Setting 1, the LSZ plug-in estimator will be biased. Since in Setting 3 the pairs (F_{12}, F_{13}) and (F_{21}, F_{23}) do not have the same entries, the LSZ normalized and Phelan estimators will be biased. On the other hand, $Q_{hj}(\tau_h) = Q_{hj}(\infty)$ when $(h, j) = (1, 3)$ and $(h, j) = (2, 1)$, and thus the LSZ plug-in estimator will perform well for P_{13} and P_{21}.

4.2 The Results

We evaluated our estimators and the existing estimators for the generated data in each simulation scenario. The simulation study was based on 1000 repetitions.

4.2.1 Estimation of transition probability—In each scenario, we evaluated the LSZ plug-in estimator $\hat{P}_{hj}$, the LSZ normalized estimator $\tilde{P}_{hj}$, and the Phelan estimator $\check{P}_{hj}$. The sample means of the LSZ plug-in estimators are close to the true values for $c_{max} = 5$ but not for $c_{max} = 3$. The sample means of the LSZ normalized estimator and the Phelan estimator are close to the true values when assumption (4) is satisfied, and they show observable bias when the assumption is violated regardless of the censoring time. The detailed results are available upon request.

We constructed 95% confidence intervals for the transition probabilities by the nonparametric bootstrap approach (see Section 2.3) with the bootstrap sample size $B = 500$. We also constructed 95% confidence intervals using the three existing estimators in the three simulation settings, assuming that the estimators are all consistent. For Setting 3, Table 1 presents the coverage frequencies and the sample mean lengths for the approximate confidence intervals. A summary of the confidence intervals for Settings 1 and 2 is available upon request.

As expected, the coverage of the confidence intervals based on the existing methods is rather low when the corresponding point estimates are biased, especially when $c_{max} = 3$. In contrast, the confidence intervals based on our approach contain the attainable values of the

transition probabilities $[P_{hj}^L, P_{hj}^U]$ at approximately the nominal level, and thus they cover the true transition probabilities at the nominal level or higher. This verifies the robustness of our interval estimator to the violation of assumption (4).

On the other hand, we observe that the confidence intervals based on our approach are relatively wide, especially when $c_{max} = 3$. They are comparable in length with the interval estimates based on the existing approaches when $c_{max} = 5$. For example, in Setting 3 with $n = 100$ and $c_{max} = 5$, the sample mean length (ML) of our interval for the transition probability P_{12} is 0.25, while the MLs of the confidence intervals based on the LSZ plug-in and normalized estimators and the Phelan estimator are 0.24, 0.23, and 0.22, respectively.

In general, the MLs for our approach are longer than the corresponding MLs for the existing estimators, assuming that they are consistent. As discussed in Section 2.3, this is because the proposed confidence interval is constructed to cover at the nominal level the interval $[P_{hj}^L, P_{hj}^U]$, which includes the true P_{hj}, and the probability limits of the existing estimators, P_{hj}^L for the LSZ plug-in estimator and $Q_{hj}(\tau_h) / \Sigma_{k \neq h} Q_{hk}(\tau_h)$ for the LSZ normalized estimator.

4.2.2 Estimation of semi-Markov kernel and sojourn time distribution—For the simulated data in each setting, we also constructed confidence bands for the semi- Markov kernel and for the attainable sojourn time distributions $[F_{hj}^L(\cdot), F_{hj}^U(\cdot)]$. We used both the EP and the HW weights, and we considered both direct construction and construction with the double-log transformation.

For illustration purposes, the domains of the confidence bands were restricted to lie within [0.5,1.5]. For Setting 3, the coverage frequencies of confidence bands with a nominal level of 95% are summarized in Table 2. Those without the transformation are slightly lower than the nominal level, and those with the transformation are close to this level, especially for the small sample size, $n = 50$. The improvement brought about by the transformation is more substantial for the EP bands than the HW bands. This is similar to findings in the context of classical survival analysis, where this transformation gives a substantial improvement in performance for the confidence intervals and confidence bands of a survival function with the Kaplan–Meier estimator (Borgan and Liestol, 1990).

5 Application to Hospitalization Data

The hospitalization data mentioned in Section 1 were collected during 1986–2000 from a group of 1374 cancer survivors who were diagnosed before the age of 20 in British Columbia from 1981 to 1995 and had survived five years or longer after diagnosis on entry into the study. The primary goal of the CAYACS study was to assess the long-term resource needs of childhood cancer survivors and to develop strategies to improve access to and the effectiveness of medical care; seeMcBride et al. (2010). The hospitalization records of each subject are available from his/her study entry to his/her death or December 31, 2000. Viewing hospitalization as a recurrent event,Hu et al. (2011) analyze the data by modifying commonly used recurrent-event approaches to address the non-ignorable event duration resulting from a hospital stay. They focus on assessing the effects of pre-identified factors.

This section presents an analysis of the hospitalization data using the multi-state process framework considered in this paper. Cook and Lawless (2007) suggest modelling a hospitalization process as an alternating two-state process with the states “in hospital” and “out of hospital”. We used a three-state semi-Markov process with state 1 for “health”, 2 for “in hospital”, and 3 for “death”. We viewed the available hospitalization records as

independent realizations of the process subject to right-censoring. We assumed that the censoring was noninformative. Our modelling avoids the potential informative censoring caused by death under the alternating two-state process model. In addition, the formulation allows us to make inferences on the two important transitions: “health” to “death” and “in hospital” to “death”. There were 60 deaths, 29 of which occurred when the subjects were “in hospital”. We aimed to estimate the probabilities of transitions between different states and the sojourn time distributions associated with transitions. This may provide important insights into the dynamics of the hospitalization process.

The “death” state is absorbing, and thus the associated kernel functions $Q_{31}(\cdot) = Q_{32}(\cdot) = 0$ and the transition probabilities $P_{31} = P_{32} = 0$. Table 3 presents estimates of the nontrivial transition probabilities using the four approaches discussed in this paper. All the approaches gave similar estimates and confidence intervals for P_{21} and P_{23}. In fact, the estimates of P_{21} and P_{23} are the empirical proportions since no subject’s data collection was censored when he/she was “in hospital”. This was mostly because of the relatively short hospital duration (i.e., short sojourn times in state 2 relative to the other sojourn times). Another cause could be a delay in the reporting of hospitalizations, which could lead to informative censoring. There is only a small difference between the hospitalization data and the corresponding portion of the hospitalization data collected until 2004 by the CAYACS team. This indicates that reporting delays did not lead to many missing hospitalization records. Therefore, the noninformative censoring assumption is plausible for this data set.

The estimates of P_{12} and P_{13} are quite different. The LSZ plug-in estimates give $\hat{P_{12}} + \hat{P_{13}} = 0.784 < 1$ because of the relatively heavy censoring. Of the 1374 subjects, the observations for 810 were censored without any transition recorded. The LSZ normalized estimates $\tilde{P_{12}}$ and $\tilde{P_{13}}$ are close to the corresponding Phelan estimates $\check{P}_{12}$ and $\check{P}_{13}$. However, they may be biased since it is unlikely that the distributions of the sojourn times in the “health” state are the same regardless of the next state, which is either “in hospital” or “death”. The possible violation of assumption (4) is shown by Figure 1, which presents the estimates and confidence bands for the sojourn time distributions starting from “health”; these results are obtained from a preliminary analysis using the LSZ normalized estimator. The construction in general gives valid confidence bands, provided assumption (4) holds. However, the confidence bands in Figure 1 indicate an observable difference between distributions $F_{12}(\cdot)$ and $F_{13}(\cdot)$, suggesting that assumption (4) is violated. This explains the discrepancy between the LSZ normalized and the Phelan estimates on one hand, and the estimates obtained by the proposed approach. In fact, Figure 1 indicates that $\hat{F_{12}}(\cdot) \leq \hat{F_{13}}(\cdot)$, i.e., the transtion from “health” to “death” is shorter than that to “in hospital”. This is rather counterintuitive. It provides strong motivation for using our method to estimate the transition probabilities P_{12} and P_{13} and the sojourn time distributions. However, as expected, the confidence intervals for P_{12} and P_{13} are rather wide because of the heavy censoring associated with “health”.

In Figure 2, we present the confidence bands for the semi-Markov kernel for the hospitalization process. The transformed HW and EP bands appear similar. The $Q_{21}(\cdot)$ and $Q_{23}(\cdot)$ estimates in Figure 2 indicate that about 90% of the subjects in hospital survive and are discharged within 15 days, while about 1% of the subjects admitted to hospital die at the hospital within a month. The $Q_{12}(\cdot)$ and $Q_{13}(\cdot)$ estimates show that about 50% of subjects survive more than 2.5 years without hospitalization, and less than 1% of the subjects die out of hospital within a year of discharge without further hospitalization.

Figure 3 presents the confidence bands for the sojourn time distributions based on our approach. The band for $F_{12}(\cdot)$ is rather wide compared with that for $F_{21}(\cdot)$. This is due to the heavy censoring, which likely results in a rather wide interval $[P_{12}^L, P_{12}^U]$. The wide confidence band for $F_{13}(\cdot)$ reflects the relatively little information in the data on the

transition to death. On the other hand, the estimates of $F_{21}(\cdot)$ suggest that about 95% of the hospital durations (i.e., the sojourn time from "in hospital" to "health") are shorter than 15 days.

We used $B = 1000$ as the resampling size in the procedures for evaluating the estimators of the transition probabilities, the semi-Markov kernel functions, and the sojourn time distributions. The corresponding critical values are presented in Table 4 along with those for $B = 500$. The critical values for the two resampling sizes are very close, which indicates that the resampling procedures discussed in Section 3 are quite stable when the resampling sizes are reasonably large.

Our data analysis is tailored for the purpose of illustrating the estimation procedures developed in Sections 2 and 3. We have assumed that the hospitalization process is a homogeneous semi-Markov process, which may not be plausible. Because of this and the possible informative censoring arising from reporting delays, we advised the medical team to interpret the analysis outcomes with caution. In addition, the data were collected during 1985–2000 from young cancer survivors diagnosed from 1981 to 1995. Caution is recommended when applying our results to more general populations.

6 Discussion

The classical competing risks process is a special case of a semi-Markov process with one transient state and K absorbing states. In this simple setting,Gaynor et al. (1993) argue for the practical importance of the transition probability and the sojourn time distribution. They provide an example showing that the transition probability can be nonidentifiable and the normalized estimator can be misleading. The transition probability, also called the case fatality ratio, has been studied recently in the competing risks setting (Jewell et al., 2007). However, to the best of our knowledge, no reference in the literature has proposed inference procedures for the transition probability based on its attainable values, even in the simple competing risks setting.

The justification of our interval estimator in Section 2.3 assumed that $P(Y_h(\tau_h) > 0) > 0$ and applied the asymptotic properties of $\hat{Q}_{hj}(\cdot)$ derived by Gill (1980). There are many practical situations with $P(Y_h(\tau_h) > 0) > 0$, including the example discussed in this paper. It is challenging to analytically derive the asymptotic distribution of $(\hat{P}_{hj}^L, \hat{P}_{hj}^U)'$ when $P(Y_h(\tau_h) > 0) = 0$ (Gill, 1980; Phelan, 1990). However, in the simulation study reported in Section 4, the nonparametric bootstrap approach worked well when $P(Y_h(\tau_h) > 0) = 0$. A similar bootstrap method has been successfully used to evaluate the performance of confidence intervals for the transition probability based on the plug-in and normalized estimators in the simple competing risks setting (Jewell et al., 2007).

This paper considers estimation based on right-censored homogeneous semi-Markov (HSM) processes, in which the transition intensities depend on both the present state and its duration. The homogeneity assumption may not hold in some practical situations where the transitions may also be associated with other history information for the process. An extension of the HSM model is the modulated semi-Markov model (Cox, 1973), which handles nonhomogeneity by incorporating a time-dependent covariate in the Cox regression form. A further generalization, the nonhomogeneous semi-Markov process (Iosifescu Manu, 1972), assumes that its transition intensity involves two time scales, the individual study time since the onset of the process and the duration in the current state. Estimation with the two more general models may be used to check the HSM assumption.

The proposed estimation procedures require that the censoring time is independent of the multi-state process. This can be a problem in some applications. We could extend the current approaches to incorporate covariates and thus handle dependent censoring that becomes independent conditional on some covariates. It would also be of interest to develop methods to deal with other informative censoring. A further investigation would be to explore adaptations of the proposed approaches to situations with different data structures, such as the panel data studied in Kang and Lagakos (2007).

References

Andersen, PK.; Borgan, O.; Gill, RD.; Keiding, N. Statistical Models Based on Counting Processes. New York: Springer-Verlag; 1993.

Andersen PK, Esbjerg S, Sorensen TA. Multi-state models for bleeding episodes and mortality in liver cirrhosis. Statistics in Medicine. 2000; 19:587–599. [PubMed: 10694738]

Andersen PK, Keiding N. Multi-state models for event history analysis. Statistical Methods in Medical Research. 2002; 11:91–115. [PubMed: 12040698]

Borgan Ø, Liestøl K. A note on confidence intervals and bands for the survival function based on transformations. Scandinavian Journal of Statistics. 1990; 17:35–41.

Commenges D. Multi-state models in epidemiology. Lifetime Data Analysis. 1999; 5:315–327. [PubMed: 10650740]

Cook, RJ.; Lawless, JF. The Statistical Analysis of Recurrent Events. New York: Springer; 2007.

Cox, DR. The statistical analysis of dependencies in point processes. In: Lewis, PAW., editor. Symposium on Point Processes. New York: Wiley; 1973. p. 55-66.

Dinse GE, Larson MG. A note on semi-Markov models for partially censored data. Biometrika. 1986; 73:379–386.

Gaynor JJ, Feuer EJ, Tan CC, Wu DH, Little CR, Straus DJ, Clarkson BD, Brennan MF. On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. Journal of the American Statistical Association. 1993; 88:400–409.

Gill RD. Nonparametric estimation based on censored observations of a Markov renewal process. Z. Wahrsch. verw. Gebiete. 1980; 53:97–116.

Hall WJ, Wellner JA. Confidence bands for a survival curve from censored data. Biometrika. 1980; 67:133–143.

Hu XJ, Lorenzi M, Spinelli J, Ying S, McBride M. Analysis of recurrent events with non-negligible event duration, with application to assessing hospital utilization. Lifetime Data Analysis. 2011; 17:215–233. [PubMed: 20730625]

Iosifescu Manu A. Non homogeneous semi-Markov processes. Studii si Cercetuari Matematice. 1972; 24:529–533.

Jewell NP, Lei X, Ghani AC, Donnelly CA, Leung GM, Ho LM, Cowling BJ, Hedley AJ. Non-parametric estimation of the case fatality ratio with competing risks data: an application to Severe Acute Respiratory Syndrome (SARS). Statistics in Medicine. 2007; 26:1982–1998. [PubMed: 16981181]

Kang M, Lagakos SW. Statistical methods for panel data from a semi-Markov process, with application to HPV. Biostatistics. 2007; 8:252–264. [PubMed: 16740624]

Lagakos SW, Sommer C, Zelen M. Semi-Markov models for partially censored data. Biometrika. 1978; 65:311–317.

Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika. 1993; 80:557–572.

Matthews DE. Some observations on semi-Markov models for partially censored data. The Canadian Journal of Statistics. 1984; 12:201–205.

McBride ML, Rogers PC, Sheps SB, Glickman V, Broemeling AM, Goddard K, Hu XJ, Lorenzi M, Peacock S, Pritchard S, Rassekh SR, Siegel L, Spinelli JJ, Teckle P, Xie L. Childhood, adolescent, young adult cancer survivors research program of British Columbia: Objectives, study design and cohort characteristics. Pediatric Blood Cancer. 2010; 55(2):324–330. [PubMed: 20582971]

Nair VN. Confidence bands for survival functions with censored data: A comparative study. Technometrics. 1984; 26:265–275.

Phelan MJ. Estimating the transition probabilities from censored Markov renewal processes. Statistics & Probability Letters. 1990; 10:43–47.

Pollard, D. Empirical Processes: Theory and Applications. Regional Conference Series in Probability and Statistics. Vol. 2. Hayward, CA: Institute of Mathematical Statistics; 1990.

Ross, SM. Stochastic Processes. Second Edition. New York: Wiley; 1996.

Appendix: Theoretical Derivation

We outline below the proofs of the propositions in Sections 2 and 3.

A.1. Proof of Proposition 1

Since $Y_h^n(u)$ is a left-continuous function of u, we have that sup $\{u:Y_h^n(u)>0\}$ = max $\{u:Y_h^n(u)>0\}$, which is the largest observed or censored sojourn time in state h, denoted by $V_h^{(n)}$. Note that $\hat{Q}_{hj}(\cdot)$ does not change after $V_h^{(n)}$. Thus,

$$P\left(\widehat{Q}_{hj}(V_h^{(n)})=\widehat{Q}_{hj}(\infty)\right)=1. \quad (16)$$

We first show that $P(\hat{Q}_{hj}(\tau_h) = \hat{Q}_{hj}(\infty)) = 1$. By the definition of τ_h, for any $\tau' > \tau_h$, we have $P(Y_h(\tau') = 0) = 1$. Thus,

$$P(Y_h^n(\tau')=0)=\prod_{i=1}^{n}P(Y_{hi}(\tau')=0)=1.$$

By the definition of $V_h^{(n)}$, $P\left(V_h^{(n)}<\tau'\right) \geq P(Y_h^n(\tau')=0)$. Thus, $P(V_h^{(n)}<\tau')=1$. It follows that

$$P(V_h^{(n)} \leq \tau_h)=1. \quad (17)$$

Combining (16) and (17) with the monotonicity of $\hat{Q}_{hj}(\cdot)$, we have $P(\hat{Q}_{hj}(\tau_h) = \hat{Q}_{hj}(\infty)) = 1$.

Further, by the continuity of $Q_{hj}(\cdot)$ and Theorem 1 of Gill (1980), $\hat{Q}_{hj}(\tau_h)$ converges to $Q_{hj}(\tau_h)$ in probability. It follows that the LSZ plug-in estimator of P_{hj}, $\hat{P}_{hj} = \hat{Q}_{hj}(\infty)$, converges to $Q_{hj}(\tau_h)$ in probability. Similarly, the LSZ normalized estimator P_{hj} converges to $Q_{hj}(\tau_h)/\sum_{k\neq h} Q_{hk}(\tau_h)$ in probability. This proves the proposition.

A.2. Proof of Proposition 2

It follows by Theorem 3 of Gill (1980) that $\{n^{1/2}\{\hat{Q}_{hj}(\tau) - Q_{hj}(\tau)\}: \tau \in [0, \nu_h]\}$ converges weakly to a Gaussian process with a covariance function, denoted by $\Sigma_{hj}(\tau_1, \tau_2)$. In addition, according to Gill (1980), on the domain $\{\tau:Y_h^n(\tau)>0$ and $1 - H_h(\tau-) > 0\}$, it can be shown by integration by parts that

$$n^{1/2}\{\widehat{Q}_{hj}(\tau)-Q_{hj}(\tau)\}=n^{-1/2}\int_0^{\tau}(1-\widehat{H}_h(u-))\frac{n}{Y_h^n(u)}dZ_{hj}^n(u)+n^{-1/2}\int_0^{\tau-}[Q_{hj}(u)-Q_{hj}(\tau)]\frac{1-\widehat{H}_h(u-)}{1-H_h(u)}\frac{n}{Y_h^n(u)}dZ_h^n(u) \quad (18)$$

where

$$Z_{hji}(u)=N_{hji}(u)-\int_0^u Y_{hi}(v)\frac{dQ_{hj}(v)}{1-H_h(v-)},$$

$Z_{hj}^n(u)=\sum_{i=1}^n Z_{hji}(u)$, and $Z_h^n(u)=\sum_{j\in\mathscr{E}} Z_{hj}^n(u)$. By the uniform law of large numbers (Pollard, 1990, page 39), $\hat{H}_h(\cdot)$ is uniformly consistent for $H_h(\cdot)$, and $Y_h^n(\cdot)/n$ is uniformly consistent for $y_h(\cdot) = E(Y_h(\cdot))$. It follows that $n^{1/2}\{\hat{Q}_{hj}(\tau) - Q_{hj}(\tau)\}$ given in (18) has the same asymptotic distribution as $n^{-1/2}\sum_{i=1}^n A_{hji}(\tau)$, a sum of n independent and identically distributed terms, where

$$A_{hji}(\tau)=\int_0^\tau \frac{1-H_h(u)}{y_h(u)}dZ_{hji}(u)+\int_0^{\tau-}\frac{Q_{hj}(u)-Q_{hj}(\tau)}{y_h(u)}d\sum_{k\in\mathscr{E}} Z_{hki}^n(u), \quad (19)$$

Thus, $\Sigma_{hj}(\tau_1, \tau_2)$ can be consistently estimated by $\frac{1}{n}\sum_{i=1}^n A_{hji}(\tau_1)A_{hji}(\tau_2)$.

The quantity $W_{hj}^n(\tau)$ in (9) is attained by replacing $Z_{hji}(\cdot)$, $Q_{hj}(\cdot)$, and $H_h(\cdot)$ on the righthand side of (18) with $\hat{Z}_{hji}(\cdot)$ U_i, $\hat{Q}_{hj}(\cdot)$, and $\hat{H}_h(\cdot)$, respectively, where $\{U_i: i = 1,\ldots, n\}$ are independent standard normal random variables that are independent of the available data, and

$$\hat{Z}_{hji}(u)=N_{hji}(u)-\int_0^u Y_{hi}(v)\frac{d\hat{Q}_{hj}(v)}{1-\hat{H}_h(v-)}=N_{hji}(u)-\int_0^u Y_{hi}(v)\frac{dN_{hj}^n(v)}{Y_h^n(v)}.$$

Note that conditional on the data, only $\{U_i: i = 1,\ldots, n\}$ are random. By the functional central limit theorem (Pollard, 1990, page 51), $\{W_{hj}^n(\tau):\tau\in[0,\nu_h]\}$ converges weakly to a Gaussian process. It can easily be shown that its covariance function has the same limit as $\frac{1}{n}\sum_{i=1}^n A_{hji}(\tau_1)A_{hji}(\tau_2)$. Thus, $\{W_{hj}^n(\tau):\tau\in[0,\nu_h]\}$ and $\{n^{1/2}\{\hat{Q}_{hj}(\tau) - Q_{hj}(\tau)\}:\tau\in[0, \nu_h]\}$ converge weakly to the same Gaussian process.

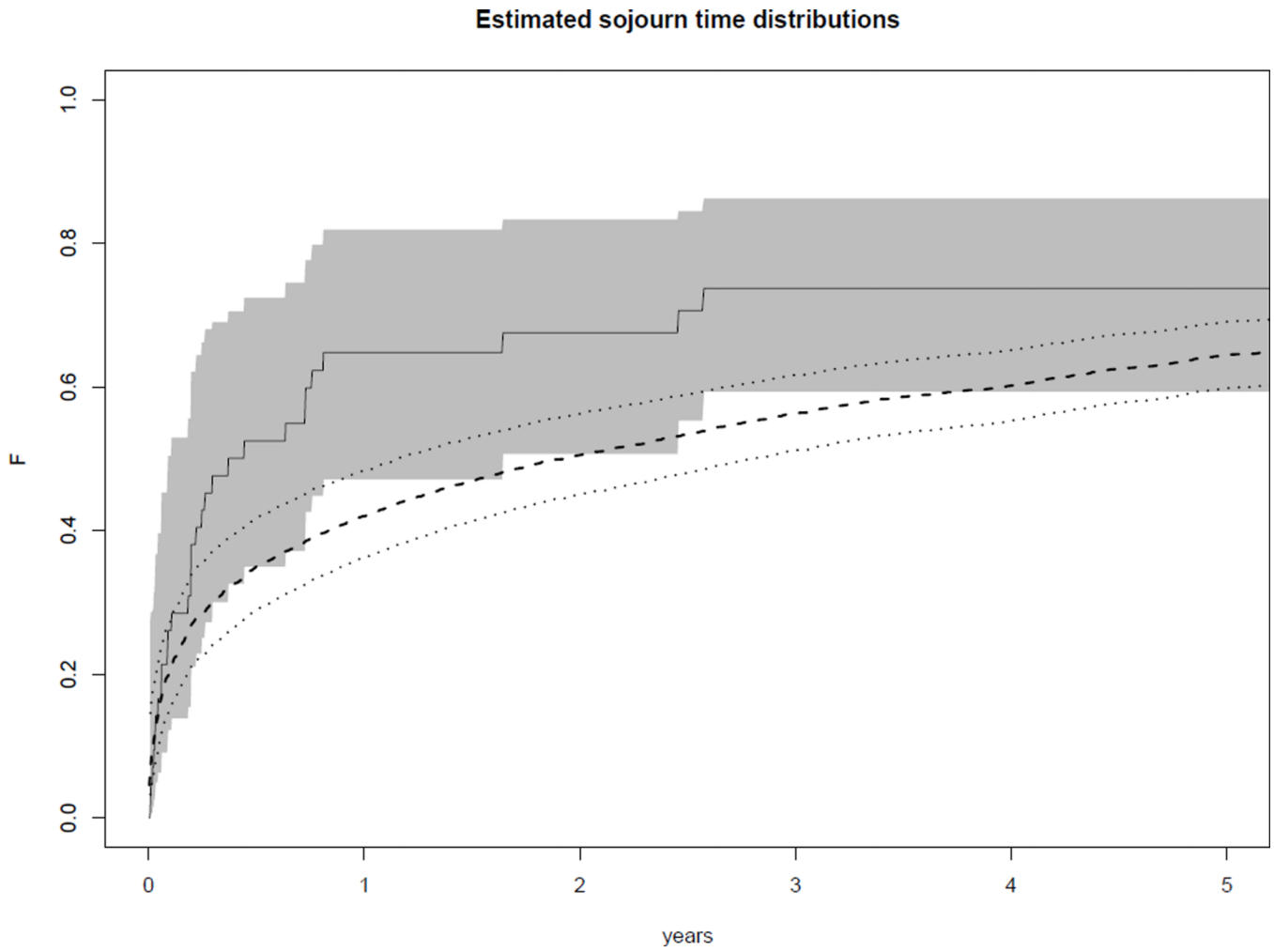

Figure 1.
Estimated sojourn time distributions for the hospitalization data based on the LSZ normalized estimator. (Dashed: point estimate of F_{12}; dotted: 95% confidence bands for F_{12}; solid: point estimate of F_{13}; shaded: 95% confidence bands for F_{13}.)

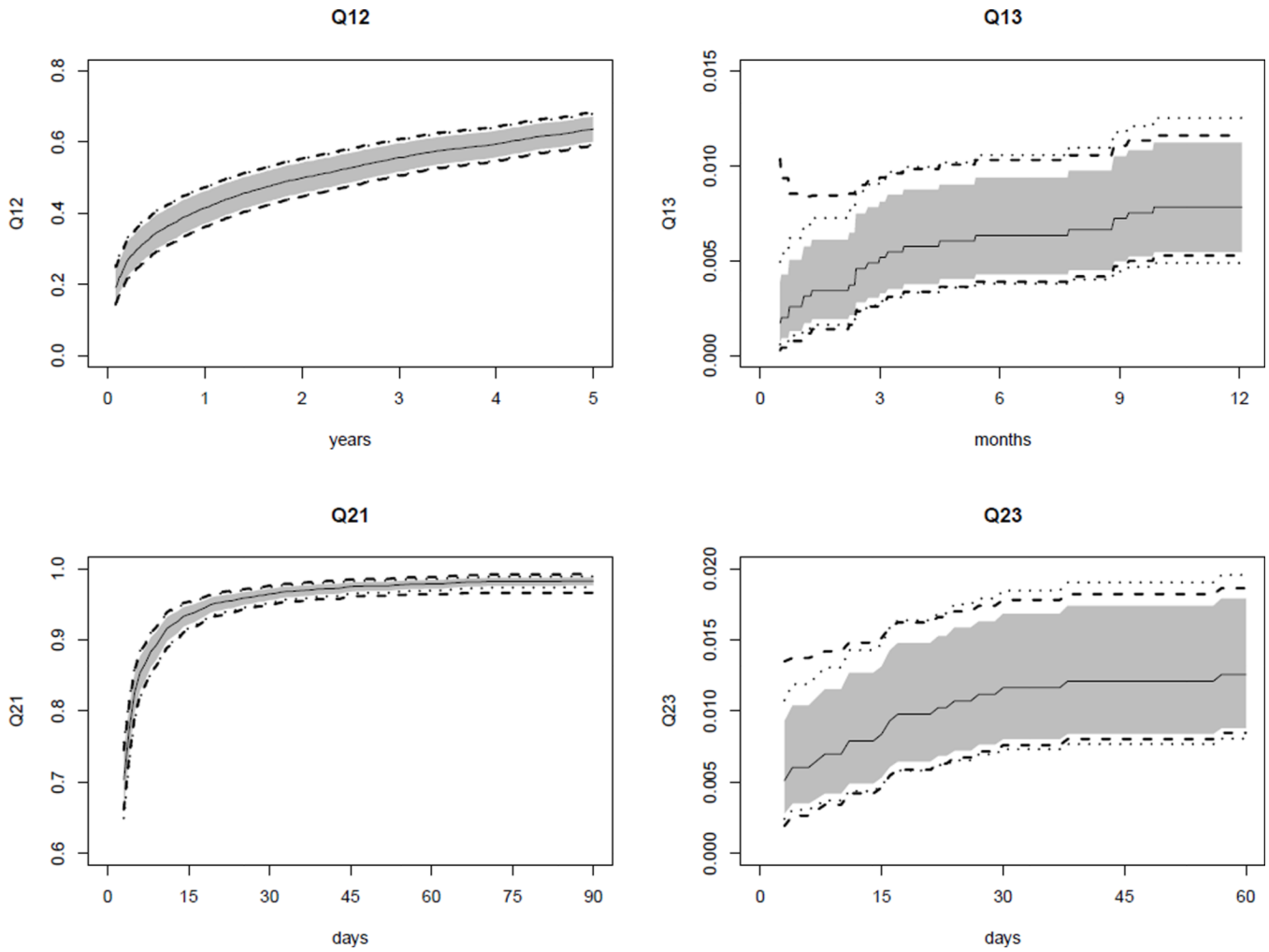

Figure 2.
Semi-Markov kernel estimates for the hospitalization data. (Solid: point estimates; shaded: 95% pointwise confidence intervals; dotted: 95% transformed EP bands; dashed: 95% transformed HW bands.)

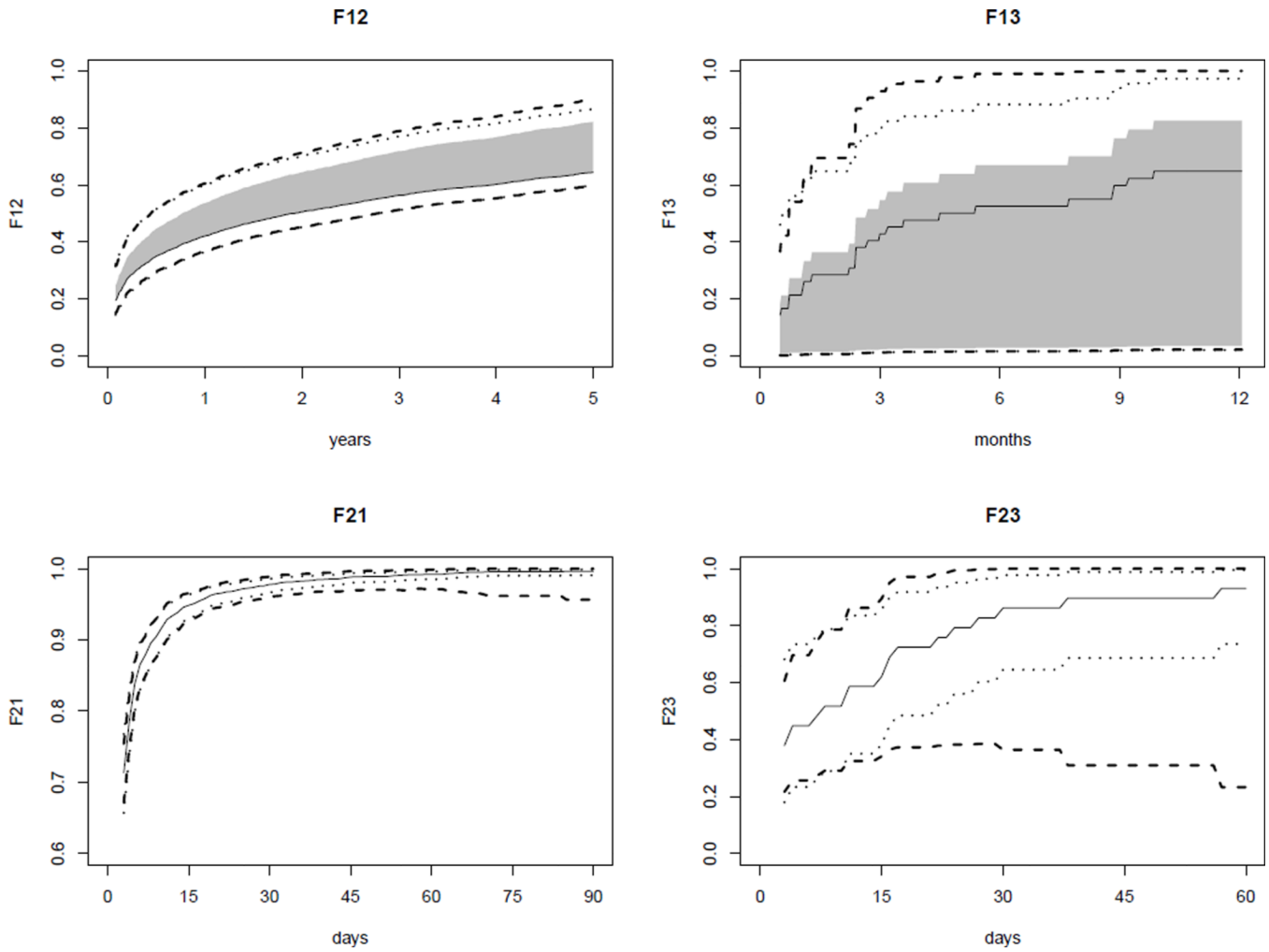

Figure 3.
Estimates of sojourn time distributions for the hospitalization data. (Solid: LSZ normalized estimates; shaded: estimated bounds ($\widehat{F}_{hj}^{L}(\cdot), \widehat{F}_{hj}^{U}(\cdot)$); dotted: 95% transformed EP bands; dashed: 95% transformed HW bands.)

Table 1

Empirical coverage probabilities (CP) and sample mean lengths (ML) of the 95% confidence intervals for the transition probabilities of the embedded Markov chain based on 1000 simulations and Setting 3

		LSZ Plug-in		LSZ Normalized		Phelan		Proposed Approach		
n	c_{max}	CP	ML	CP	ML	CP	ML	CP1[1]	CP2[2]	ML
$P_{12} = 0.7$										
50	3	0.88	0.42	0.91	0.41	0.92	0.37	0.96	0.95	0.47
	5	0.93	0.33	0.94	0.33	0.94	0.31	0.96	0.95	0.35
100	3	0.86	0.31	0.92	0.29	0.91	0.26	0.97	0.95	0.35
	5	0.94	0.24	0.94	0.23	0.93	0.22	0.96	0.96	0.25
200	3	0.87	0.22	0.95	0.21	0.93	0.18	0.98	0.96	0.26
	5	0.95	0.17	0.96	0.16	0.95	0.16	0.97	0.97	0.18
$P_{13} = 0.3$										
50	3	0.90	0.39	0.91	0.41	0.92	0.37	0.96	0.95	0.47
	5	0.94	0.32	0.94	0.33	0.94	0.31	0.96	0.95	0.35
100	3	0.90	0.28	0.92	0.29	0.91	0.26	0.97	0.95	0.35
	5	0.94	0.22	0.94	0.23	0.93	0.22	0.96	0.96	0.25
200	3	0.95	0.20	0.95	0.21	0.93	0.18	0.98	0.96	0.26
	5	0.96	0.16	0.96	0.16	0.95	0.16	0.97	0.97	0.18
$P_{21} = 0.5$										
50	3	0.92	0.43	0.82	0.45	0.84	0.41	0.96	0.95	0.57
	5	0.94	0.34	0.92	0.36	0.88	0.34	0.98	0.97	0.42
100	3	0.92	0.31	0.81	0.32	0.78	0.29	0.97	0.95	0.44
	5	0.94	0.24	0.92	0.26	0.82	0.24	0.97	0.96	0.31
200	3	0.94	0.22	0.74	0.24	0.67	0.21	0.98	0.96	0.35
	5	0.94	0.17	0.91	0.18	0.70	0.17	0.97	0.96	0.24
$P_{23} = 0.5$										
50	3	0.65	0.41	0.82	0.45	0.84	0.41	0.96	0.95	0.57
	5	0.85	0.36	0.92	0.36	0.88	0.34	0.98	0.97	0.42

		LSZ Plug-in		LSZ Normalized		Phelan		Proposed Approach		
n	c_{max}	**CP**	**ML**	**CP**	**ML**	**CP**	**ML**	**CP1**[1]	**CP2**[2]	**ML**
100	3	0.57	0.32	0.81	0.32	0.78	0.29	0.97	0.95	0.44
	5	0.83	0.27	0.92	0.26	0.82	0.24	0.97	0.96	0.31
200	3	0.46	0.24	0.74	0.24	0.67	0.21	0.98	0.96	0.35
	5	0.79	0.20	0.91	0.18	0.70	0.17	0.97	0.96	0.24

[1] CP for the transition probabilities

[2] CP for the attainable intervals of the transition probabilities

Can J Stat. Author manuscript; available in PMC 2014 June 01.

Table 2

Empirical coverage probabilities of the equal precision (EP), Hall–Wellner (HW), transformed equal precision (TEP), and transformed Hall–Wellner (THW) 95% confidence bands for the semi-Markov kernel and attainable sojourn time distributions based on 1000 simulations and Setting 3

		Semi-Markov Kernel				Sojourn Time Distribution			
n	c_{max}	EP	HW	TEP	THW	EP	HW	TEP	THW
Transition 1 →2									
50	3	0.88	0.91	0.93	0.94	0.83	0.92	0.96	0.95
	5	0.89	0.92	0.93	0.94	0.87	0.93	0.95	0.94
100	3	0.92	0.93	0.93	0.94	0.90	0.94	0.96	0.95
	5	0.90	0.92	0.93	0.93	0.90	0.94	0.95	0.95
200	3	0.93	0.93	0.93	0.94	0.92	0.93	0.94	0.94
	5	0.93	0.94	0.94	0.94	0.93	0.95	0.94	0.95
Transition 1 →3									
50	3	0.77	0.89	0.94	0.92	0.78	0.90	0.93	0.92
	5	0.81	0.90	0.92	0.91	0.78	0.83	0.97	0.94
100	3	0.83	0.90	0.93	0.92	0.86	0.89	0.96	0.96
	5	0.86	0.93	0.93	0.92	0.85	0.87	0.95	0.96
200	3	0.87	0.91	0.94	0.94	0.91	0.90	0.96	0.96
	5	0.89	0.92	0.92	0.93	0.90	0.91	0.94	0.94
Transition 2 →1									
50	3	0.80	0.89	0.93	0.92	0.80	0.88	0.93	0.92
	5	0.81	0.90	0.91	0.91	0.87	0.90	0.96	0.96
100	3	0.85	0.91	0.92	0.93	0.88	0.89	0.95	0.94
	5	0.89	0.92	0.93	0.93	0.90	0.91	0.95	0.95
200	3	0.90	0.92	0.93	0.93	0.92	0.93	0.94	0.94
	5	0.91	0.93	0.93	0.93	0.92	0.94	0.94	0.94
Transition 2 →3									
50	3	0.85	0.92	0.93	0.92	0.84	0.91	0.93	0.92
	5	0.84	0.90	0.92	0.92	0.87	0.90	0.94	0.94
100	3	0.85	0.91	0.92	0.92	0.91	0.91	0.96	0.96

		Semi-Markov Kernel				Sojourn Time Distribution			
n	c_{max}	EP	HW	TEP	THW	EP	HW	TEP	THW
200	5	0.88	0.92	0.93	0.92	0.90	0.92	0.94	0.94
	3	0.90	0.91	0.94	0.94	0.93	0.93	0.96	0.96
	5	0.92	0.94	0.94	0.93	0.93	0.94	0.94	0.94

Table 3

Point estimates and 95% confidence intervals for the transition probabilities of the hospitalization data (state 1: “health”, 2: “in hospital”, 3: “death”)

		LSZ Plug-in	LSZ Normalized	Phelan	Proposed approach
P_{12}	Estimate	0.775	0.988	0.986	0.775–0.991
	CI	(0.738, 0.813)	(0.984, 0.992)	(0.981, 0.991)	(0.738, 0.994)
P_{13}	Estimate	0.009	0.012	0.014	0.009–0.225
	CI	(0.006, 0.013)	(0.008, 0.016)	(0.009, 0.019)	(0.006, 0.262)
P_{21}	Estimate	0.986	0.986	0.986	0.986–0.986
	CI	(0.982, 0.991)	(0.982, 0.991)	(0.982, 0.991)	(0.982, 0.991)
P_{23}	Estimate	0.014	0.014	0.014	0.014–0.014
	CI	(0.009, 0.018)	(0.009, 0.018)	(0.009, 0.018)	(0.009, 0.018)

Table 4

Critical values determined via two different resampling sizes to evaluate the estimates in the hospitalization data analysis

	Transition Probability		Semi-Markov Kernel $q_{hj}(s_1, s_2)$		Sojourn Time Distribution $q_{hj}^L(s_1, s_2) = q_{hj}^U(s_1, s_2)$	
B	c_1	c_2	**EP**	**HW**	**EP**	**HW**
Transition 1 →2 ($[s_1, s_2] = [30, 1825]$)						
500	0.038	0.003	2.45	1.13	2.47	1.08
1000	0.037	0.003	2.44	1.13	2.49	1.08
Transition 1 →3 ($[s_1, s_2] = [15, 365]$)						
500	0.003	0.038	2.54	0.11	2.71	0.59
1000	0.003	0.037	2.56	0.11	2.69	0.58
Transition 2 →1 ($[s_1, s_2] = [3, 90]$)						
500	0.005	0.005	2.68	0.75	2.80	0.74
1000	0.005	0.005	2.72	0.74	2.83	0.72
Transition 2 →3 ($[s_1, s_2] = [3, 60]$)						
500	0.005	0.005	2.50	0.18	2.65	0.35
1000	0.005	0.005	2.45	0.18	2.70	0.35