

Non-Parametric Inference of Transition Probabilities Based on Aalen-Johansen Integral Estimators for Acyclic Multi-State Models: Application to LTC Insurance

Quentin Guibert, Frédéric Planchet

▶ **To cite this version:**

Quentin Guibert, Frédéric Planchet. Non-Parametric Inference of Transition Probabilities Based on Aalen-Johansen Integral Estimators for Acyclic Multi-State Models: Application to LTC Insurance. 2018. hal-01183542v2

HAL Id: hal-01183542
https://hal.science/hal-01183542v2

Preprint submitted on 11 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Non-Parametric Inference of Transition Probabilities Based on Aalen-Johansen Integral Estimators for Acyclic Multi-State Models: Application to LTC Insurance

Quentin Guibert*[1,2] and Frédéric Planchet†[1,2]

[1]Univ Lyon, Université Claude Bernard Lyon 1, Institut de Science Financière et d'Assurances (ISFA), Laboratoire SAF EA2429, F-69366, Lyon, France
[2]Prim'Act, 42 avenue de la Grande Armée, 75017 Paris, France

May 24, 2018

Abstract

Studying Long Term Care (LTC) insurance requires modeling the lifetime of individuals in presence of both terminal and non-terminal events which are concurrent. Although a non-homogeneous semi-Markov multi-state model is probably the best candidate for this purpose, most of the current researches assume, maybe abusively, that the Markov assumption is satisfied when fitting the model. In this context, using the Aalen-Johansen estimators for transition probabilities can induce bias, which can be important when the Markov assumption is strongly unstated. Based on some recent studies developing non-Markov estimators in the illness-death model that we can easily extend to a more general acyclic multi-state model, we exhibit three non-parametric estimators of transition probabilities of paying cash-flows, which are of interest when pricing or reserving LTC guarantees in discrete time. As our method directly estimates these quantities instead of transition intensities, it is possible to derive asymptotic results for these probabilities under non-dependent random right-censorship, obtained by re-setting the system with two competing risk blocks. Inclusion of left-truncation is also considered. We conduct simulations to compare the performance of our transition probabilities estimators without the Markov assumption. Finally, we propose a numerical application with LTC insurance data, which is traditionally analyzed with a semi-Markov model.

Keywords: Multi-state model; Aalen-Johansen integral; non-parametric estimator; non-Markov process; LTC insurance.

*Email: q.guibert@hotmail.fr.
†Email: frederic.planchet@univ-lyon1.fr.

1 Introduction

Multi-state models offer a sound modeling framework for the random pattern of states experienced by an individual along time. These stochastic models are very flexible and can be adapted to many applications. In biostatistics (Hougaard, 1999; Hougaard, 2001; Andersen and Keiding, 2002), this specification is generally used to model the transitions between states, defined as the occurrence of a disease or a serious event affecting the survival of an individual. For credit risk and reliability areas, this framework is transposed to account for the lifetime history of a firm or an item, see e.g. Lando and Skødeberg (2002) and Janssen and Manca (2007). For fifteen years, multi-state models have provoked a growing interest in the actuarial literature modeling the random pattern of states experienced by a policyholder during the contract period. In this context, transitions between states occur when an event triggers the payment of premiums and benefits. For health and life insurance modeling purposes, many papers develop comprehensive frameworks for pricing and reserving both with Markov or semi-Markov assumptions (Haberman and Pitacco, 1998; Denuit and Robert, 2007). Christiansen (2012) gives a wide overview of the use of multi-state models in health insurance, including Long Term Care (LTC) insurance, from an academic perspective. For that purpose, actuaries need to estimate the transition probabilities between states and additionally the transition intensities if a continuous underlying model is used. In practice, these probabilities may be adjusted to account for complex policy conditions (e.g. waiting periods and deferral periods).

Fitting multi-state models related to disability and LTC insurance with the available data is generally done with regression approaches, in a manner similar to mortality models. Most approaches in the empirical literature resort to the Markov assumption, i.e. the transition to the next state depends only on the current state, see e.g. Gauzère et al. (1999), Pritchard (2006), Deléglise et al. (2009), Levantesi and Menzietti (2012), Fong et al. (2015) among others. This allows to keep calculations simple, but the process ignores the effects of the past lifetime-path. This assumption is also often used, since the available data for fitting the model are rare with few fine details. However, it is clearly inappropriate when modeling the LTC claimants mortality, as the transition probabilities depend on the occurring age and the duration (or sojourn time) of each disease, and a semi-Markov model seems to be more relevant. In the actuarial literature, research about fitting non-Markov models is relatively scarce and focused mainly on disability data, which are generally fitted with parametric models, e.g. the so-called Poisson model (Haberman and Pitacco, 1998) or the Cox semi-Markov model (Czado and Rudolph, 2002). For a semi-Markov without any loop, the most common approach consists in estimating each crude transition intensity, as the ratio between the number of transitions from one state to another and the exposure at risk. Then, a Poisson regression model is applied on both one-dimensional and two-dimensional multiple decrements tables, along similar lines to the smoothing approaches developed for one or two-dimensional mortality tables, see e.g. Currie et al. (2004) on the use of generalized additive models or generalized linear models for this purpose. As noted by Tomas and Planchet (2013), the LTC claimants mortality rates have a complex pattern that requires using flexible smoothing techniques such as p-splines or local methods. Recently for LTC insurance contracts, Biessy (2015) and Fuino and Wagner (2017) use semi-Markov models with a Weibull law for the duration time in disability states, which can be easily implemented and are quite flexible. The calculation of the transition probabilities is carried out in a third step by solving Kolmogorov differential equations based on the smoothed transition intensities.

In this paper, we focus on multi-state model with both multiple terminal (e.g. multiple causes of death) and non-terminal (e.g. competing diseases or degrees of disability) events without possibility of recovery, which is adapted to many LTC insurance specifications. This type of multi-state model contains at most two jumps, in connection with the future cash-flows of the contract, and can be represented by a semi-Markov process. However when estimating the model, the usual approaches

described above can fail as several assumptions are violated. First, the usual Markov assumption is likely to be wrong in disability states. Second, the semi-Markov structure can be questioned depending on how the data are observed. In an estimation framework with disability or LTC insurance data, the Markov assumption is actually lost when a policyholder become disabled, e.g. Alzheimer's disease or other degenerative diseases. This entry date is generally unobserved as the insurer only records the entry date in a dependency state as settled contractually. As a consequence, the observed duration can differ from the sojourn time with a chronic pathology depending on how quick the disease is diagnosed. Thus, the estimated transition probability from the healthy state to this contractual disability state could be biased if the transition intensities of a semi-Markov model are calculated on the left censored duration times. Similar issues may appear when the multi-state process is affected by exogenous effects, e.g. unobserved heterogeneity. Finally, the crude intensities used in regression approaches are calculated using a discrete-time method as the ratio between the number of transitions and the exposure. For transitions from a disability state, this requires to define carefully a (entry age, age)-diagram or a (duration, age)-diagram. In most cases, yearly or monthly death rates are considered, but for LTC guarantees a finer timescale may be necessary for the first year after the entry into dependency. This is typically what happens for disabled insured suffering from a terminal cancer, as their monthly death rates just after the entry into dependency is around 30%. This required to use diagrams with different timescales leading to several steps when smoothing the crude rates, which is awkward.

To avoid these issues, we propose a direct non-parametric estimation framework with no Markov assumption focusing on transition probabilities as key quantities for actuaries. The terminology "direct" means that this method does not require to estimate the transition intensities in a first step. Definition of these key transition probabilities depends on the terms of the policy and they are specifically exhibited to compute, in discrete time, the price or the amount of reserve related to insurance liabilities. Considering targets adapted to a non-homogeneous semi-Markov specification in line with the contract clauses, our method gives relevant estimators for these probabilities with asymptotic properties, even when the Markov assumption is not satisfied. This avoid to introduce bias when specified a Markov or semi-Markov framework based on transition intensities, and allows the construction of confidence intervals for transition probabilities, which is not possible after implementing numerical techniques for the resolution of Kolmogorov differential equations. An additional feature of our approach is that it is not needed to specify a discretization grid as our method is adapted to continuous-time data, or to introduce a parametric assumption. From a practical point of view, our estimators can be used to perform model checking, e.g. analyze the goodness of fit of a regression model, in a manner similar to that used for checking a parametric model for the survival function with the Kaplan-Meier estimator for the construction of biometric tables in insurance. In that sense, we underline that our key quantities are not smoothed.

Our approach is based on recent alternatives to the canonical Aalen-Johansen estimator for transition probabilities, which is adapted to Markov multi-state models, which have been developed for some particular non-Markov models. For a progressive (or acyclic) illness-death model, Meira-Machado et al. (2006) propose direct non-parametric estimators based on a Kaplan-Meier integral representation. Estimators exhibited by Allignol et al. (2014) are very similar, but use competing risks techniques and allow for left-truncation. This first generation of estimators has been recently criticized as they are systematically biased if the support of the observed lifetime distribution of an individual is not contained within that of the censoring distribution. Alternatives are given by de Uña-Álvarez and Meira-Machado (2015) and Titman (2015) in a progressive illness-death model, as well as some other configurations. However, our aim differs from estimating usual transition probabilities $\mathbb{P}(X_t = j \mid X_s = h)$, where X_t denotes the state of the insured at time t, as is generally the practice in biostatistical literature. Thus, we apply a model featuring two competing risk

blocks which are nested to account for the progressive form of the process with right-censored data. With such a structure, our model can be viewed as a particular case of a bivariate competing risks data problem with only one censoring process. This structure requires considering Aalen-Johansen integrals for competing risks data (Suzukawa, 2002), instead of the Meira-Machado et al. (2006)' Kaplan-Meier integral estimators. This allows to construct a first class of transition probabilities estimators that we enrich secondly with more efficient alternatives, following de Uña-Álvarez and Meira-Machado (2015). As insurance data are generally subject to left-truncation and right-censoring, this paper examines also how to adapt all these estimators in that context.

This paper is organized as follows. Section 2 introduces the modeling framework adapted to LTC insurance and defines transition probabilities for the rest of the paper. After defining Aalen-Johansen integral estimators, we derive in Section 3 three versions of the non-parametric estimators of the quantities under study. Their asymptotic properties are discussed, as well as the inclusion of left-truncation. Section 4 is devoted to a simulation analysis to assess the performance of our non-parametric estimators. We also assess the bias which appears when estimating a semi-Markov model based on data simulated with censored duration times. Application to real French LTC insurance data is proposed in Section 5. The supplementary material describes in more details the underlying estimation framework and presents the asymptotic results of our estimators, as well as some additional simulation results.

2 Multi-state model for LTC insurance

We present a semi-Markov model for a LTC insurance contract in Section 2.1 with the aim of introducing key probabilities for actuarial purposes. The interest of focusing on a direct estimation procedure for these indicators is discussed in Section 2.2, in particular regarding to the loss of the semi-Markov assumption.

2.1 A semi-Markov model for LTC insurance payments

Long Term Care (LTC) insurance is a mix of social care and health care provided on a daily basis, formally or informally, at home or in institutions, to people suffering from a loss of mobility and autonomy in their activities of daily life. In France for example, this guarantee is dedicated to elderly people who are partially or totally dependent and benefits are mainly paid as an annuity. Their amounts depend on the policyholders' lifetime-paths and possibly on their degree of dependency (see e.g. Plisson, 2009; Courbage and Roudaut, 2011).

As for classical guarantees the payments defined by the contract depend on the pattern of health states experienced by the policyholder and the sojourn time in each state, we introduce in this section a particular semi-Markov model to describe his current state, see e.g. Christiansen (2012) or Buchardt et al. (2014) for a more general presentation on semi-Markov models in similar situations for life and health insurance. This framework is fairly general and allows considering most of LTC insurance specifications. The semi-Markov structure is a natural choice as the Markov assumption is generally not satisfied for LTC insurance contracts, see e.g. Denuit and Robert (2007) and Tomas and Planchet (2013). Using the payments process, our aim here is exclusively to exhibit transition probabilities which are key quantities to compute beyond the prospective reserves as the expected present value of the cash flows.

For the rest of the paper, we employ a particular multi-state structure with no recovery that we call *acyclic*. On a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, we consider a time-continuous stochastic process $(X_t)_{t\geqslant 0}$ with finite state space $\mathcal{S} = \{a_0, e_1, \ldots, e_{m_1}, d_1, \ldots, d_{m_2}\}$, as defined by the contract, and right-continuous paths with left-hand limits. This process represents the state of the policyholder at time $t \geqslant 0$. The set $\{e_1, \ldots, e_{m_1}\}$ contains m_1 intermediary states or non-terminal events that

we associate to disability competing causes. The set $\{d_1, \ldots, d_{m_2}\}$ contains terminal events, i.e. absorbing states such as straight death, lapse or death after entry in dependency. The state a_0 corresponds to the healthy state. Hence, an individual can take two types of lifetime paths depending on whether an intermediate event occurs or not. Figure 1 depicts an example of such an acyclical multi-state structure with two levels of state considered in the LTC insurance contract.

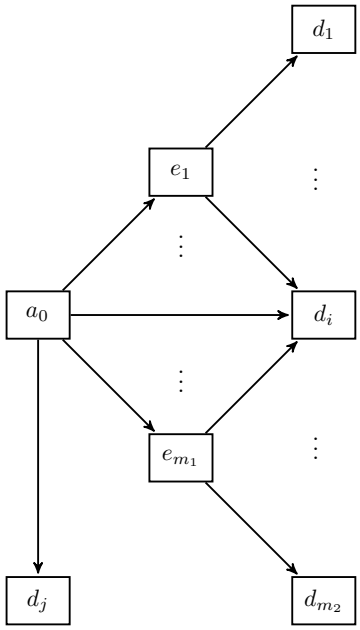


Figure 1: Example of an acyclic multi-state model with intermediary and terminal states.

We denote by $(U_t)_{t\geqslant 0}$ the sojourn time spent in the current state

$$U_t = \max\{\tau \in [0,t] : X_u = X_t, u \in [t-\tau, t]\} \ .$$

In such a context, it is natural to assume that the process $(X_t, U_t)_{t\geqslant 0}$ satisfies the Markov assumption. The process $(X_t)_{t\geqslant 0}$ is then called a semi-Markov process and is characterized by the transition probabilities, for all $0 \leqslant u \leqslant s \leqslant t$, $v \geqslant 0$ and $h, j \in \mathcal{S}$

$$p_{hj}(s,t,u,v) = \mathbb{P}\left(X_t = j, U_t \leqslant v \,|(X_s, U_s) = (h,u)\right). \tag{2.1}$$

We also define the transition intensities for all $0 \leqslant u \leqslant t$ and $h, j \in \mathcal{S}$,

$$\mu_{hj}(t,u) = \lim_{\delta t \to 0} \frac{p_{hj}(t, t+\delta t, u, \infty)}{\delta t} \text{ for } h \neq j \text{ and } \mu_{hh}(t,u) = -\sum_{j \neq h} \mu_{hj}(t,u) \ , \tag{2.2}$$

which we suppose exist and are continuous. Note in particular that any acyclic multi-state model as in Figure 1 is a semi-Markov model, if X is endogenously generated. Finally, we introduce $\bar{p}_{hh}(s,t,u)$ as the probability to not stay in state h within $[s,t]$ with duration u in time s and use the letter Δ to indicate that the sojourn time is included into a time interval such as

$$p_{hj}(s,t,\Delta_u,\Delta_v) = \mathbb{P}\left(X_t = j, v-1 < U_t \leqslant v \mid X_s = h, u-1 < U_s \leqslant u\right) \ ,$$

and

$$\bar{p}_{hh}(s,t,\Delta_u) = \mathbb{P}\left(X_t \neq h \mid X_s = h, u-1 < U_s \leqslant u\right).$$

Under this semi-Markov framework , we can derive premiums and reserves expressions related to a LTC insurance policy. For that, we recall the general formula of the expected present value of benefits, net of premiums, paid within a period $[t, \infty[$ for a policyholder initially in state $h \in \{a_0, e_1, \ldots, e_{m_1}\}$ with duration u

$$\begin{aligned} & V_h(t,u) \\ & = \int_t^\infty \delta(t,\tau) \sum_{j \in \mathcal{S}} \int_0^{\tau-t+u} p_{hj}(t,\tau,u,dv) \left(dB_j(\tau,v) + \sum_{l \neq j} \mu_{jl}(\tau,v)\, b_{jl}(\tau,v) \right) d\tau, \end{aligned} \tag{2.3}$$

where $\delta(t,\tau) = e^{-\int_t^\tau r_s ds}$, r_t denotes the continuous compounded risk free interest rate, $dB_j(t,U_t) = b_j(t,U_t)\,dt$ is a continuous, net of premiums, annuity rate payments accumulated in state j during the sojourn time U_t to the policyholder at time t. $b_{jl}(t,U_t)$ is the single payment related to the transition from state j to state l. The payments and the interest rate functions are assumed to be continuous and deterministic. Of course, when a insured reaches a state $\{d_1, \ldots, d_{m_2}\}$, prospective reserves are nil after the payment of the last benefits. Depending upon contractual conditions, the net benefits can easily include deferred periods and a stopping time.

Equation (2.3) can be differentiated with respect to t and u as a generalized Thiele's differential equation (Hoem, 1972; Denuit and Robert, 2007). A common used method to derive prospective reserves with the Thiele equation consists in calculating the transition probabilities, given that the transition rates are specified, with the so-called Kolmogorov's backward or Kolmogorov's forward differential equations (see e.g. Buchardt et al., 2014).

In our particular case where the multi-state structure does not admit any loop (or any return to a previous state) and only two jumps, computing procedures are simplified. In particular, $p_{hh}(t,\tau,u,dv)$, $h \in \{a_0, e_1, \ldots, e_{m_1}\}$ are zero for $v \neq \tau + u - t$, otherwise $p_{hh}(t,\tau,u,dv) = 1 - \bar{p}_{hh}(t,\tau,u)$. Since the states $d_1 \ldots, d_{m_2}$ are terminal states, a payment may occur only at the transition time. Thus, we get

$$\begin{aligned} V_{a_0}(t,u) = & \int_t^\infty \delta(t,\tau)\left(1 - \bar{p}_{a_0a_0}(t,\tau,u)\right) dB_{a_0}(\tau,\tau-t+u)\, d\tau \\ & + \int_t^\infty \delta(t,\tau)\left(1 - \bar{p}_{a_0a_0}(t,\tau,u)\right) \sum_{l \neq a_0} \mu_{a_0l}(\tau,\tau-t+u)\, b_{a_0l}(\tau,\tau-t+u) d\tau \\ & + \int_t^\infty \delta(t,\tau) \sum_{j=e_1}^{e_{m_1}} \int_0^{\tau-t} p_{a_0j}(t,\tau,u,dv) \left(dB_j(\tau,v) + \sum_{l \neq j} \mu_{jl}(\tau,v)\, b_{jl}(\tau,v) \right) d\tau, \end{aligned} \tag{2.4}$$

and, for $e \in \{e_1, \ldots, e_{m_1}\}$

$$\begin{aligned} V_e(t,u) = & \int_t^\infty \delta(t,\tau)\left(1 - \bar{p}_{ee}(t,\tau,u)\right) dB_e(\tau,\tau-t+u)\, d\tau \\ & + \int_t^\infty \delta(t,\tau)\left(1 - \bar{p}_{ee}(t,\tau,u)\right) \sum_{l=d_1}^{d_{m_2}} \mu_{el}(\tau,\tau-t+u)\, b_{el}(\tau,\tau-t+u) d\tau. \end{aligned} \tag{2.5}$$

In many practical situations, a discrete time approach is chosen to derive the prospective reserves, by assuming that transitions only happen at integer times (only one transition per period) and payments are null at non-integer times. For simplicity and without a significant loss of generality, we suppose the payments and transitional probabilities from the state a_0 do not depend on the duration u, i.e. the Markov assumption is verified for the state a_0. Thus, Equation (2.4) is rewritten

in discrete time (with a time-scale of 1) with payments in advance and a sojourn time in state a_0 equal to 0

$$\begin{aligned} V_{a_0}(t,0) &= \sum_{\tau=t}^{\infty} \delta(t,\tau)\left(1-\bar{p}_{a_0a_0}(t,\tau,0)\right)\left(B_{a_0}(\tau)-B_{a_0}(\tau-1)\right) \\ &+ \sum_{\tau=t}^{\infty} \delta(t,\tau+1)\left(1-\bar{p}_{a_0a_0}(t,\tau,0)\right)\sum_{l\neq a_0} p_{a_0l}(\tau,\tau+1,0,\infty)\, b_{a_0l}(\tau+1) \\ &+ \sum_{\tau=t+1}^{\infty} \delta(t,\tau)\sum_{j=e_1}^{e_{m_1}}\sum_{v=1}^{\tau-t} p_{a_0j}(t,\tau,0,\Delta_v)\left(B_j(\tau,v)-B_j(\tau-1,v-1)\right) \\ &+ \sum_{\tau=t+1}^{\infty} \delta(t,\tau+1)\sum_{j=e_1}^{e_{m_1}}\sum_{v=1}^{\tau-t}\sum_{l\neq j} q_{a_0jl}(t,\tau,0,\Delta_v)\, b_{jl}(\tau,v)\,, \end{aligned} \tag{2.6}$$

where

$$p_{a_0j}(t,\tau,0,\Delta_v) = \mathbb{P}\left(X_\tau = j, v-1 < U_\tau \leqslant v \mid X_t = a_0\right),$$

is the transition probability from state a_0 to state j between time t and time τ and with a sojourn time into $\Delta_v =]v-1, v]$, and

$$q_{a_0jl}(t,\tau,0,\Delta_v) = \mathbb{P}\left(X_\tau = j, X_{\tau+1} = l, v-1 < U_\tau \leqslant v \mid X_t = a_0\right),$$

is the transition probability from state a_0 to state l between time t and time τ and with a sojourn time in an intermediary state j between $v-1$ and v [1]. Similarly, Equation (2.5) becomes for $e \in \{e_1, \ldots, e_{m_1}\}$

$$\begin{aligned} V_e(t,u) &= \sum_{\tau=t}^{\infty} \delta(t,\tau)\left(1-\bar{p}_{ee}(t,\tau,\Delta_u)\right)\left(B_e(\tau,\tau-t+u)-B_e(\tau-1,\tau-1-t+u)\right) \\ &+ \sum_{\tau=t}^{\infty} \delta(t,\tau+1)\left(1-\bar{p}_{ee}(t,\tau,\Delta_u)\right)\sum_{l=d_1}^{d_{m_2}} p_{el}(\tau,\tau+1,\Delta_{\tau-t+u},\infty)\, b_{el}(\tau+1,\tau+1-t+u). \end{aligned} \tag{2.7}$$

With the above discretized prospective reserves, we are interested in estimating probabilities of paying (or receiving) cash-flows:

- **i.** $\bar{p}_{a_0a_0}(s,t,0)$, $0 \leqslant s \leqslant t$,
- **ii.** $p_{a_0j}(s,t,0,\infty)$, for a state $j \neq a_0$ and $0 \leqslant s \leqslant t$,
- **iii.** $p_{a_0e}(s,t,0,\Delta_v)$, for a non-terminal state due to cause e, $0 \leqslant s \leqslant t$ and a sojourn time into $\Delta_v =]v-1, v]$, $v \geqslant 0$,
- **iv.** $q_{a_0ed}(s,t,0,\Delta_v)$, for a non-terminal state due to cause e, a terminal cause d, $0 \leqslant s \leqslant t$ and a sojourn time in state e within $\Delta_v =]v-1, v]$, $v \geqslant 0$,
- **v.** $\bar{p}_{ee}(s,t,\Delta_u)$, with a non-terminal event e and $0 \leqslant u \leqslant s \leqslant t$,
- **vi.** $p_{ed}(s,t,\Delta_u,\infty)$, for a non-terminal event e, a terminal event d and $0 \leqslant u \leqslant s \leqslant t$.

[1] For the sake of simplicity, we use the same time interval to discretize the process along the time and the duration dimensions.

2.2 Discussion on the estimation approach

For the semi-Markov framework introduced above, the transition probabilities could be easily carried out by solving Kolmogorov differential equations in continuous time based on the estimates of transition intensities. The latter are usually fitted by means of a Poisson regression model or a parametric model (e.g. Weibull duration laws). In this paper, we have chosen to estimate directly these probabilities, i.e. without using the transition intensities, and without any Markov assumption. Several motivations are discussed below.

About the semi-Markov assumption As the definition of the state space $\mathcal{S}$ is in line with the contract clauses, the use of the semi-Markov model X might introduce bias in practical situations. In the LTC insurance context where a insured is exposed to progressive illnesses, such as Alzheimer's disease, it is reasonable to assume that his health is influenced by several intermediary states (see e.g. Salazar et al., 2007). In particular, the sojourn time in each intermediary state may be a good candidate for explaining the evolution of the health state. However, these states could remain unobserved from the insurer's point of view as individual data collected by insurance companies are incomplete and ignore usually the medical information. This situation is illustrated in Figure 2 with the introduction of latent states influencing the current health state of the insured.

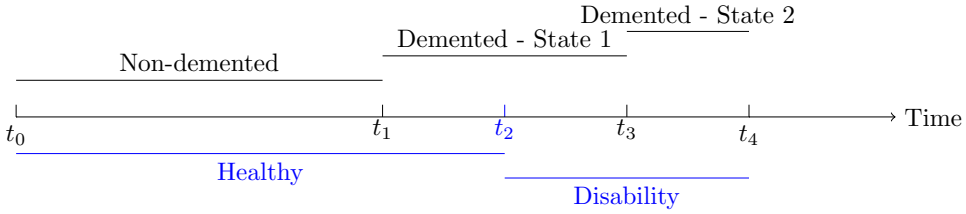


Figure 2: Example of follow-up of one insured. The blue lines represent the time spend in each observed state, as defined by the contract. The black lines is for the sojourn time in some intermediary states related to a cognitive disease, e.g. dementia.

In this example, the insured is seen healthy by the insurer at time t_1 and the claim is reported at time t_2, whereas the insured is actually demented for a duration time of $t_2 - t_1$. In addition, the insurer ignores the transition to a more severe dementia state at time t_3, which may affect the occurrence of the death at time t_4. In such a situation, the duration time in the disability state is probably not enough to capture the dynamic of the process. Hence, a semi-Markov multi-state model based on an extended state space would be more relevant than a model based on contractual states. However, this extended state space is unattainable since the transition times into the intermediary states are unobserved. Our non-Markov approach is one way to solve this problem, as shown in Section 4.2.

An alternative procedure for model checking The robustness of usual approaches depends on the quality of the transition intensities defined in two dimensions. Both for a parametric model and a Poisson regression model, this requires to compute crude intensity rates for analyzing the goodness-of-fits. These rates are based on a piecewise constant hypothesis on a duration×age rectangle or on a duration×entry age parallelogram. With individual data, the both versions are available. Estimating the crude intensity rates requires to define carefully the timescales used for aggregating along two dimensions the number of transitions and the contributions to the exposure at risk. In most cases, yearly or monthly rates are considered, but as an expert opinion. Our approach avoid to choose an arbitrary timescale, which is appealing for model checking. In addition, our estimators

allow to focus directly on the transition probabilities of interest and not on the transition intensities which can be seen as intermediary outcomes. Another argument is that our approach allows us to derive asymptotic properties for the estimates, which is not possible when estimating the transition intensities and then resolving the Kolmogorov differential equations.

3 Non-parametric estimation of transition probabilities

This section outlines in 3.1 our notations and explains the link between transition probabilities of interest and a bivariate competing risks structure for estimation purposes. We do this in a heuristic fashion for right-censoring data and give more details regarding the underlying framework, the asymptotic properties of the estimators and the related proofs in the supplementary material. In particular, this material discusses the case where discrete covariates are added. Section 3.2 investigates alternative estimators to the first approach avoiding systematic biases and improves efficiency. Proofs are also available in the supplementary material. As LTC insurance data are also subject to left-truncation, we ultimately explain in Section 3.3 how to adapt our estimators to comply with it.

3.1 Estimation with Aalen-Johansen integrals

In case the process is Markovian, transition probabilities of interest do not depend on the sojourn time in a disability state and can be estimated non-parametrically with the so-called Aalen-Johansen estimator (Aalen and Johansen, 1978; Andersen et al., 1993). However, this methodology fails when the Markov assumption is wrong, especially when these probabilities depend on both time and duration (semi-Markov models). In what follows, we develop a non-parametric approach to estimate directly these transition probabilities, instead of calculating transition intensities.

The proposed estimation approach for acyclic multi-state models is based on two competing risks processes which are nested. In a first step, the individual lifetime history can be affected by competing exit-causes from the healthy state, i.e. to the non-terminal ($a_0 \to e$) and to the terminal ($a_0 \to d$) states. In a second step, the residual lifetime is exposed to m_2-causes of exit, may they be direct ($a \to d$) or indirect through one of the m_1 intermediary states ($a \to e \to d$).

Let (S, V_1) and (T, V) be two competing risk models, where S is the sojourn time in the healthy state a_0, V_1 indicates the exit cause of the state a_0, T is the overall survival time and $V = (V_1, V_2)$ with V_2 indicating the reached absorbing state. As noted by Meira-Machado et al. (2006) with an acyclic illness-death model, $S = T$ and V_1 is identical to V_2 when a direct transition occurs initially. Otherwise, we have $S < T$ and V_1 codes necessarily an intermediary state. Note also that V_2 depends on the value taken by V_1, in accordance with the multi-state structure.

The data used for estimation are subject to censoring and the two lifetime variables (S, T) are not directly observed. We consider a right censoring variable C, which is assumed to be independent to the vector (S, T, V). This assumption is widely used for simplicity in practice and is generally verified by insurance data as observations are censored by administrative events or simply due to the end of the observation period. It is important to note that the censoring variable is unique. Thus, the following variables are available

$$\begin{cases} Y = \min(S, C) \text{ and } \gamma = \mathbb{1}_{\{S \leqslant C\}}, \\ Z = \min(T, C) \text{ and } \delta = \mathbb{1}_{\{T \leqslant C\}}. \end{cases}$$

The observation of the i-th individual of a sample of length $n \geqslant 1$ is characterized by

$$(Y_i, \gamma_i, \gamma_i V_{1,i}, Z_i, \delta_i, \delta_i V_{2,i}) \;\; 1 \leqslant i \leqslant n \;,$$

which are assumed to be i.i.d. replications of the variable $(Y, \gamma, \gamma V_1, Z, \delta, \delta V_2)$. If $\delta = 1$, then obviously $\gamma = 1$. Thus, the transition probabilities of interest can be expressed in terms of the joint distributions (S, V_1) and (S, T, V) as follows

$$\bar{p}_{a_0 a_0}(s,t,0) = \frac{\mathbb{P}(s < S \leqslant t)}{\mathbb{P}(S > s)}, \tag{3.1}$$

$$p_{a_0 j}(s,t,0,\infty) = \frac{\mathbb{P}(s < S \leqslant t, V_1 = j)}{\mathbb{P}(S > s)}, \tag{3.2}$$

$$p_{a_0 e}(s,t,0,\Delta_v) = \frac{\mathbb{P}(s < S \leqslant t < T, t - S \in \Delta_v, V_1 = e)}{\mathbb{P}(S > s)}, \tag{3.3}$$

$$q_{a_0 ed}(s,t,0,\Delta_v) = \frac{\mathbb{P}(s < S \leqslant t < T \leqslant t+1, t - S \in \Delta_v, V = (e,d))}{\mathbb{P}(S > s)}, \tag{3.4}$$

$$\bar{p}_{ee}(s,t,\Delta_u) = \frac{\mathbb{P}(S \leqslant s < T \leqslant t, s - S \in \Delta_u, V_1 = e)}{\mathbb{P}(S \leqslant s < T, s - S \in \Delta_u, V_1 = e)}, \tag{3.5}$$

$$p_{ed}(s,t,\Delta_u,\infty) = \frac{\mathbb{P}(S \leqslant s < T \leqslant t, s - S \in \Delta_u, V = (e,d))}{\mathbb{P}(S \leqslant s < T, s - S \in \Delta_u, V_1 = e)}. \tag{3.6}$$

Even if the Markov assumption is not verified, the probability (3.1) can be simply estimated using the Kaplan-Meier estimator of the distribution function of S, noted H and assumed to be continuous

$$\widehat{H}_n(s) = \sum_{i=1}^{n} W_{in} \mathbb{1}_{\{Y_{i:n} \leqslant s\}}, \tag{3.7}$$

where $Y_{1:n} \leqslant Y_{2:n} \leqslant \ldots \leqslant Y_{n:n}$ is the ordered Y-values, $\gamma_{[i:n]}$ the concomitant of the i-th order statistic (i.e. the value of $(\gamma_j)_{1 \leqslant j \leqslant n}$ paired with $Y_{i:n}$) and

$$W_{in} = \frac{\gamma_{[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\gamma_{[j:n]}},$$

the Kaplan-Meier weight for the the i-th ordered observation. An estimator of the numerator in Equation (3.2) is obtained considering the Aalen-Johansen estimator for competing risks data of the cumulative function $H^{(e)}(s) = \mathbb{P}(S \leqslant s, V_1 = e)$, i.e.

$$\widehat{H}_n^{(e)}(s) = \sum_{i=1}^{n} W_{in}^{(e)} \mathbb{1}_{\{Y_{i:n} \leqslant s\}}, \tag{3.8}$$

where $W_{in}^{(e)} = W_{in} \mathbb{1}_{\{V_{1,[i:n]} = e\}}$.

Other probabilities include more complex terms which depend on (S, T, V). By defining the bivariate cumulative incidence function (assumed to be continuous)

$$F^{(e,d)}(s,t) = \mathbb{P}(S \leqslant s, T \leqslant t, V = (e,d)), \ s \leqslant t,$$

these complex terms can be expressed as integrals of the form $\int \varphi \, dF^{(e,d)}$, where φ is a simple transformation of S and T. This bivariate cumulative incidence function is defined and estimated nonparametrically by Cheng et al. (2007) under independent right-censoring. However, their representation is devoted to general bivariate competing risks data and we aim to provide estimators which exploit the information that S is necessarily observed when T is not censored. For

this purpose, it is convenient to introduce $Z_{1:n} \leqslant Z_{2:n} \leqslant \ldots \leqslant Z_{n:n}$ the ordered Z-values and $\left(Y_{[i:n]}, \delta_{[i:n]}, J^{(e,d)}_{[i:n]}\right)$ the concomitant of the i-th order statistic with $J^{(e,d)}_i = \mathbb{1}_{\{V_i=(e,d)\}}$.

Based on the idea of Meira-Machado et al. (2006), we consider S as a covariate and estimate $F^{(e,d)}$ using the so-called Aalen-Johansen estimator (Aalen and Johansen, 1978) adapted to a single competing risks model. This estimator can be written as

$$\begin{aligned}\widehat{F}^{(e,d)}_n(y,z) &= \sum_{i=1}^{n} \widetilde{W}^{(e,d)}_{in} \mathbb{1}_{\left\{Y_{[i:n]} \leqslant y, Z_{i:n} \leqslant z\right\}} \\ &= \sum_{i=1}^{n} \widetilde{W}_{in} J^{(e,d)}_{[i:n]} \mathbb{1}_{\left\{Y_{[i:n]} \leqslant y, Z_{i:n} \leqslant z\right\}},\end{aligned} \tag{3.9}$$

where $\widetilde{W}_{in}$ denotes the Kaplan-Meier weight of the i-th ordered observation, related to the estimated survival function of T. The Aalen-Johansen weights for states (e,d), defined as

$$\widetilde{W}^{(e,d)}_{in} = \frac{\delta_{[i:n]} J^{(e,d)}_{[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}}, \; 1 \leqslant i \leqslant n,$$

are very close to the Kaplan-Meier weights related to the estimated survival function of T. They can be interpreted as the mass associated to one observation. We also denote $J^{(e)}_i = \mathbb{1}_{\{V_{1,i}=e\}}$ and $\widetilde{W}^{(e)}_{in} = \sum_{i=1}^{d_{m_2}} \widetilde{W}^{(e,d_i)}_{in}$[2].

Note that the Inverse Probability of Censoring Weighting (IPCW) representation can be easily derived from this expression by writing[3]

$$\widetilde{W}^{(e,d)}_{in} = \frac{\delta_{[i:n]} J^{(e,d)}_{[i:n]}}{n\left(1-\widehat{G}_n\left(Y_{i:n}\right)\right)},$$

where $\widehat{G}_n$ represents the Kaplan-Meier estimator of the distribution function of C. The IPCW theory is largely used for survival models with dependent censoring and was recently applied to state occupation, exit and waiting times probabilities for acyclic multi-state models (Mostajabi and Datta, 2013) and to transition probabilities for the illness-death model (Meira-Machado et al., 2014). The extension of our estimators with these regression techniques would be clearly feasible here, but is out of the scope of this paper.

Based on the representation as a sum of (3.9), we are now interested in obtaining estimators for probabilities (3.3-3.6). Since the joint distribution of (T,V) has the aspect of a competing risks model, we propose the following estimators

$$\widehat{p}_{a_0 e}(s,t,0,\Delta_v) = \frac{\sum_{i=1}^{n} \widetilde{W}^{(e)}_{in} \mathbb{1}_{\left\{s<Y_{[i:n]} \leqslant t<Z_{i:n},\, t-Y_{[i:n]} \in \Delta_v\right\}}}{1-\widehat{H}_n(s)}, \tag{3.10}$$

$$\widehat{q}_{a_0 e d}(s,t,0,\Delta_v) = \frac{\sum_{i=1}^{n} \widetilde{W}^{(e,d)}_{in} \mathbb{1}_{\left\{s<Y_{[i:n]} \leqslant t<Z_{i:n} \leqslant t+1,\, t-Y_{[i:n]} \in \Delta_v\right\}}}{1-\widehat{H}_n(s)}, \tag{3.11}$$

$$\widehat{\widehat{p}}_{ee}(s,t,\Delta_u) = \frac{\sum_{i=1}^{n} \widetilde{W}^{(e)}_{in} \mathbb{1}_{\left\{Y_{[i:n]} \leqslant s<Z_{i:n} \leqslant t,\, s-Y_{[i:n]} \in \Delta_u\right\}}}{\sum_{i=1}^{n} \widetilde{W}^{(e)}_{in} \mathbb{1}_{\left\{Y_{[i:n]} \leqslant s<Z_{i:n},\, s-Y_{[i:n]} \in \Delta_u\right\}}}, \tag{3.12}$$

[2]Of course $\{V_1=e\} = \{V_1=e, V_2 \in \{d_1,\ldots,d_{m_2}\}\} = \{V_1=e, V_2 \in \mathcal{C}(e)\}$, where $\mathcal{C}(e)$ is the set of states to which a direct transition from e is possible.

[3]If there is no tie.

$$\widehat{p}_{ed}\left(s,t,\Delta_u,\infty\right)=\frac{\sum_{i=1}^{n}\widetilde{W}_{in}^{(e,d)}\mathbb{1}_{\left\{Y_{[i:n]}\leqslant s<Z_{i:n}\leqslant t,\ s-Y_{[i:n]}\in\Delta_u\right\}}}{\sum_{i=1}^{n}\widetilde{W}_{in}^{(e)}\mathbb{1}_{\left\{Y_{[i:n]}\leqslant s<Z_{i:n},\ s-Y_{[i:n]}\in\Delta_u\right\}}}. \tag{3.13}$$

As Y is considered as an uncensored covariate when T is not censored, these estimators are consistent w.p.1, if the support of Z is included in that of C, and they verify consistency and weak convergence properties. To obtain these results with independant right-censoring, we derive general asymptotic properties for Aalen-Johansen integrals estimators with competing risks data by refining the estimators exhibited by Suzukawa (2002) with the addition of covariates.

The asymptotic variance functions of these estimated transition probabilities are tricky to infer. To get around this difficulty, bootstrap or jackknife techniques can be used. In section 5, we construct non-parametric bootstrap pointwise confidence bands for our estimators. This is done with a simple bootstrap resampling procedure (Efron, 1979). Although it seems that jackknife techniques performs well (Azarang et al., 2015), bootstrap approaches are often preferred for practical applications as their computation cost is lower, especially when the amount of data is significant. Note also that Beyersmann et al. (2013) provide wild bootstrap approach for the Aalen-Johansen estimator for competing risks data but, as it is remarked in their paper, this approach is quite close to that followed by Efron. To the best of our knowledge, no further tentative has been proposed to obtain more consistent bootstrap methodologies for cumulative intensity function or other transition probabilities.

Finally, remark that we could write probabilities (3.2) for $j\in\{d_1,\ldots,d_{m_2}\}$ as

$$\frac{\mathbb{P}\left(s<S,T\leqslant t,V_1=V_2=j\right)}{\mathbb{P}\left(S>s\right)}.$$

In this case, these probabilities would be estimated using Aalen-Johansen integrals estimators, which are consistent at time t in the support of Z. Thus, the Aalen-Johansen weights would be involved instead of the Kaplan-Meier weights related to S. Considering the estimator (3.8) is equivalent to apply right-truncation on the observations which experience non-terminal events. However, we prefer this approach as it preserves the relation $\bar{p}_{a_0a_0}\left(s,t,0\right)=\sum_{j\neq a_0}p_{a_0j}\left(s,t,0,\infty\right)$. This seems more relevant for actuarial applications.

3.2 Alternative estimators

Although they verify asymptotic properties, the main drawback of the previous estimators is that they are systematically biased if the support of Z is not contained within that of the right-censoring distribution. This situation is rather restrictive and some alternative estimators have been recently proposed. de Uña-Álvarez and Meira-Machado (2015) introduce new estimators of transitional probabilities, $\mathbb{P}\left(X_t=j\mid X_s=h\right)$ of an illness-death model, which cope with this issue. One of their estimators is a conditional Pepe-type estimator (Pepe, 1991), i.e. the difference between two Kaplan-Meier estimates. Titman (2015) gives estimators for transitional probabilities between states or sets of states which may be used with a non-progressive multi-state model for particular transitions.

Our methodology to find alternative estimators adapts the approach developed by de Uña-Álvarez and Meira-Machado (2015) to complex transition probabilities (3.3-3.6) in an acyclic multi-state model. Following these authors, two alternative estimators are exhibited. By remarking for Equations (3.4) and (3.5) that

$$\begin{aligned}q_{a_0ed}\left(s,t,0,\Delta_v\right)&=\mathbb{P}\left(X_t=e,X_{t+1}=d,v-1<U_t\leqslant v\mid X_s=a_0\right)\\&=\mathbb{P}\left(X_{t+1}=d\mid X_s=a_0,X_t=e,v-1<U_t\leqslant v\right)\mathbb{P}\left(X_t=e,v-1<U_t\leqslant v\mid X_s=a_0\right)\\&=p_{a_0e}\left(s,t,0,\Delta_v\right)p_{ed}\left(t,t+1,\Delta_v,\infty\right),\end{aligned}$$

if $t-s>v$ by construction, and $\bar{p}_{ee}(s,t,\Delta_u)=\sum_{i=1}^{m_2}p_{ed_i}(s,t,\Delta_u,\infty)$, we limit our analysis and especially focus on estimating quantities (3.3) and (3.6).

The first solution consists in rewriting $p_{a_0e}(s,t,0,\Delta_v)$ and $p_{ed}(t,t+1,\Delta_v,\infty)$

$$\widehat{p}^{*}_{a_0e}(s,t,0,\Delta_v)=\frac{\sum_{i=1}^{n}W_{in}^{(e)}\mathbb{1}_{\{s<Y_{i:n}\leqslant t,\ t-Y_{i:n}\in\Delta_v\}}-\sum_{i=1}^{n}\widetilde{W}_{in}^{(e)}\mathbb{1}_{\left\{s<Y_{[i:n]},Z_{i:n}\leqslant t,\ t-Y_{[i:n]}\in\Delta_v\right\}}}{1-\widehat{H}_n(s)}, \tag{3.14}$$

which is consistent to $p_{a_0e}(s,t,0,\Delta_v)$ and asymptotically normal, when $\max(t,t-v+1)$ is included in the support of Y (see the left term of the numerator), and

$$\widehat{p}^{*}_{ed}(s,t,\Delta_u,\infty)=\frac{\sum_{i=1}^{n}\widetilde{W}_{in}^{(e,d)}\mathbb{1}_{\left\{Y_{[i:n]}\leqslant s<Z_{i:n}\leqslant t,\ s-Y_{[i:n]}\in\Delta_u\right\}}}{\sum_{i=1}^{n}W_{in}^{(e)}\mathbb{1}_{\{Y_{i:n}\leqslant s,\ s-Y_{i:n}\in\Delta_u\}}-\sum_{i=1}^{n}\widetilde{W}_{in}^{(e)}\mathbb{1}_{\left\{Y_{[i:n]}\leqslant s,Z_{i:n}\leqslant s,\ s-Y_{[i:n]}\in\Delta_u\right\}}}, \tag{3.15}$$

which is consistent when t is included in the support of Y or when $\max(s,s-u+1)$ and t are respectively included in those of Y and Z[4].

To increase efficiency of the estimators introduced by Meira-Machado et al. (2006), Allignol et al. (2014) construct new estimators for $\mathbb{P}(X_t=j\mid X_s=h)$ by restricting the initial sample of length n to the subpopulation at risk in state h at time s. Titman (2015) and de Uña-Álvarez and Meira-Machado (2015) follow a similar restriction. In our case, we consider first the subset $\{i:Y_i>s\}$ with cardinal ${}_sn$ and estimate $p_{a_0e}(s,t,0,\Delta_v)$ as

$$\begin{aligned}\breve{p}_{a_0e}(s,t,0,\Delta_v)=&\sum_{i=1}^{{}_sn}{}_sW_{i_sn}^{(e)}\mathbb{1}_{\{{}_sY_{i:{}_sn}\leqslant t,\ t-{}_sY_{i:{}_sn}\in\Delta_v\}}\\&-\sum_{i=1}^{{}_sn}{}_s\widetilde{W}_{i_sn}^{(e)}\mathbb{1}_{\left\{Z_{i:{}_sn}\leqslant t,\ t-Y_{[i:{}_sn]}\in\Delta_v\right\}},\end{aligned} \tag{3.16}$$

where ${}_sY$ is the lifetime in the initial state for individuals who were still in this state at time s. We introduce by analogy ${}_sW_{i_sn}^{(e)}$, ${}_s\widetilde{W}_{i_sn}^{(e)}$ and ${}_sZ$.

Second, we select individuals in the subset $\{i:Y_i<s\leqslant Z_i,s-Y_i\in\Delta_u\}$ of cardinal ${}_{s,u}n$. With this selection and unambiguous notations, $p_{ed}(s,t,\Delta_u,\infty)$ can be estimated with

$$\breve{p}_{ed}(s,t,\Delta_u,\infty)=\sum_{i=1}^{{}_{s,u}n}{}_{s,u}\widetilde{W}_{i_{s,u}n}^{(e,d)}\mathbb{1}_{\left\{{}_{s,u}Z_{i:{}_{s,u}n}\leqslant t\right\}}. \tag{3.17}$$

This corresponds to the Aalen-Johansen estimators for the cumulative incidence function of a competing risks model based on a particular subsample. Note that if there is only one terminal state which can be reached from e, Equation (3.17) corresponds to the Kaplan-Meier estimator of the distribution function of T conditionally on $\{Y<s\leqslant Z,s-Y\in\Delta_u\}$.

Similarly to the initial estimators considered in Section 3.1, bootstrap resampling techniques can be used to estimate the variance of these alternative estimators.

3.3 Estimation under right-censoring and left-truncation

Insurance datasets are often subject to left-truncation (e.g. the subscription date of the contract) and right-censoring together. Up to now, the proposed estimators are defined only under right-censoring. We discuss in this Section how to address this issue.

[4]In practice, we take $s-u+1\leqslant s$ and $t-v+1\leqslant t$.

Due to the nested structure of the model, only one truncation variable L should be considered. Hence, Y and Z are observed if $Y \geqslant L$. For LTC insurance data, which is studied in Section 5, left-truncation always occurs when the individual is still in the healthy state. This situation seems to make sense for health insurance applications, as the insurer accepts only non-dependent policyholders at the subscription date. For this reason, we exclude here more complicated situations that may arise in a more general framework if the left-truncation events occur after S (see e.g. Peng and Fine, 2006, for an illness-death model).

In this manner, we refine our proposed estimators for left-truncated and right-censored data following the representation defined by Sánchez-Sellero et al. (2005). To be specific, we assume that (C, L) is independent of (S, T, V) and C is independent of L. As we have n i.i.d. observations

$$\left(Y_i, \gamma_i, \gamma_i V_{1,i}, Z_i, \delta_i, \delta_i V_{2,i}, L_i\right), i = 1, \ldots, n,$$

only if $L_i \leqslant Y_i$, we redefine respectively the Kaplan-Meier and the Aalen-Johansen weights such as

$$W_{in} := \frac{\gamma_i}{nC_n\left(Y_i\right)} \prod_{\{j:Y_j<Y_i\}} \left(1 - \frac{1}{nC_n\left(Z_j\right)}\right)^{\gamma_j},$$

and

$$\widetilde{W}_{in}^{(e,d)} := \frac{\delta_i J_i^{(e,d)}}{n\widetilde{C}_n\left(Z_i\right)} \prod_{\{j:Z_j<Z_i\}} \left(1 - \frac{1}{n\widetilde{C}_n\left(Z_i\right)}\right)^{\delta_j},$$

where $C_n\left(x\right) = n^{-1}\sum_{i=1}^n \mathbb{1}_{\{L_i \leqslant x \leqslant Y_i\}}$ and $\widetilde{C}_n\left(x\right) = n^{-1}\sum_{i=1}^n \mathbb{1}_{\{L_i \leqslant x \leqslant Z_i\}}$. Then, these weights can be replaced in Equations (3.7-3.15) to obtain new estimators which are almost surely consistent and asymptotically normal if, in addition, the largest lower bound for the support of L is lower than that of S and $\mathbb{P}\left(L \leqslant S\right) > 0$. Estimators (3.16) and (3.17) can be adjusted in a similar way, i.e. by considering the subset $\{i : Y_i > s\}$ and $\{i : Y_i < s \leqslant Z_i, s - Y_i \in \Delta_u\}$, and then the weights pertaining to these subsamples. In this manner, we need to verify than the number of individuals at risk is not nil with a positive probability. Additional information regarding the asymptotic results are given in the supplementary material.

4 Simulation results for transition probabilities

In this section, we deploy a simulation approach to assess the performance of our proposed estimators. As noted above, we focus on probabilities (3.3) and (3.6). Two sets of simulations are done, comparing our estimators between them and with a semi-Markov model.

The computations are carried out with the software `R` (R Core Team, 2017). As our model is based on competing risks models, the `R`-package `mstate` designed by De Wreede et al. (2011) and the book of Beyersmann et al. (2011) give useful initial toolkits for the development of our code. Note that the model provided by Meira-Machado et al. (2006) is also implemented in `R` (Meira-Machado and Roca-Pardinas, 2011). We also use the package `SemiMarkov` for fitting an homogeneous semi-Markov model (Król and Saint-Pierre, 2015). Our script is available upon request made to the first author.

4.1 Comparison of the proposed estimators

The aim of this section is to compare the performance of the non-parametric estimators introduced above. For that, we introduce the multi-state structure depicted in Figure 3. For the sake of simplicity and without loss of generality, we consider only one terminal event and two non-terminal

events. Let $T_{a_0e_1}$, $T_{a_0e_2}$ and T_{a_0d} be the latent failure times in healthy state a_0 corresponding to the non-terminal states $\{e_1, e_2\}$ and to the absorbing state d. We also set T_{e_1d} and T_{e_2d} the residual lifetime from each non-terminal state due to cause d.

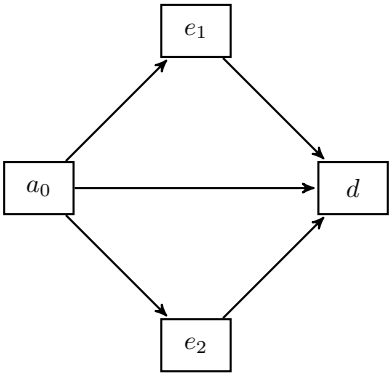


Figure 3: Multi-state structure for the simulation study.

To simulate a non-Markov process, we specify a dependence assumption between each latent failure times and consider the simulation approach set up by Rotolo et al. (2013).

First, we set a Clayton copula $\mathfrak{C}_{\theta_0}$ with dependent parameter θ_0 to combine the failure times T_{a_0e}, $e \in \{e_1, e_2\}$, and T_{a_0d} from the starting state a_0

$$\begin{aligned}\mathbb{P}\left(T_{a_0e_1} > t_{e_1}, T_{a_0e_2} > t_{e_2}, T_{a_0d} > t_d\right) &= \mathfrak{C}_{\theta_0}\left(\mathbb{P}\left(T_{a_0e_1} > t_{e_1}\right), \mathbb{P}\left(T_{a_0e_2} > t_{e_2}\right), \mathbb{P}\left(T_{a_0d} > t_d\right)\right) \\ &= \left(1 + \mathbb{P}\left(T_{a_0d} > t_d\right)^{-\theta_0} + \sum_{e \in \{e_1, e_2\}} \left[\mathbb{P}\left(T_{a_0e} > t_e\right)^{-\theta_0} - 1\right]\right)^{-1/\theta_0}.\end{aligned}$$

As a second step, we define another Clayton copula $\mathfrak{C}_{\theta_e}$ for each non-terminal event e_1, e_2 to put together its latent times T_{a_0e} and its children's T_{ed}. This gives for $e \in \{e_1, e_2\}$

$$\begin{aligned}\mathbb{P}\left(T_{a_0e} > t_e, T_{ed} > t_d\right) &= \mathfrak{C}_{\theta_e}\left(\mathbb{P}\left(T_{a_0e} > t_e\right), \mathbb{P}\left(T_{ed} > t_d\right)\right) \\ &= \left(1 + \mathbb{P}\left(T_{a_0e} > t_e\right)^{-\theta_e} + \mathbb{P}\left(T_{ed} > t_d\right)^{-\theta_e}\right)^{-1/\theta_e}.\end{aligned}$$

With this setting, we assume the dependent parameters θ_0 and θ_e, $e \in \{e_1, e_2\}$, have the same value $\theta = 0.5$ for each copula model. The latent failure times T_{a_0e}, T_{ed}, $e \in \{e_1, e_2\}$, and T_{0d} are generated with Weibull distributions $\text{Wei}\left(\lambda, \rho\right)$, where λ and ρ being respectively the scale and shape parameters. Parameters are selected in order to have features similar to that of a sample of LTC claimants. Table 1 presents the selected parameters.

To assess the performance of our estimators, we compare the results under two censoring scenarios: C follows an independent Uniform distribution $\mathcal{U}\left[30, 45\right]$ (Scenario 1), and an independent Weibull distribution $\text{Wei}\left(80, 1\right)$ (Scenario 2), i.e. an Exponential distribution. With this specification, the proportion of censoring for Scenario 1 is 14% (24% for Scenario 2) for individuals in state a_0, 33% (30% for Scenario 2) for individuals reaching the state e_1 and 38% (28% for Scenario 2) for those experiencing the state e_2. For Scenario 1, we choose the window $[30, 45]$ as many transitions to states e_1 and e_2 occur during this period. In this respect, the aim is to reveal the lack of performance of estimators introduced in Section 3.1. For each scenario, we consider three samples with size $n = 100$, $n = 200$ and $n = 400$. For $e \in \{e_1, e_2\}$, we compute $p_{a_0e}\left(s, s+4, 0, \Delta_v\right)$ and $p_{ed}\left(s, s+4, \Delta_u, \infty\right)$ at different time points: s takes the values $\tau_{.20}$, $\tau_{.40}$ and $\tau_{.60}$ corresponding to

Table 1: Simulation parameters used for the failure latent times.

Parameter	$T_{a_0e_1}$	$T_{a_0e_2}$	T_{a_0d}	T_{e_1d}	T_{e_2d}
Scale λ	35	35	50	2	5
Shape ρ	8	8	3	0.3	3

Note: This table displays the simulation parameters used with Weibull distributions $\text{Wei}(\lambda, \rho)$.

the quantiles 20%, 40% and 60% of T; and time intervals Δ_u and Δ_v both correspond to $]0, 2]$ and $]2, 4]$[5].

For both censoring scenarios, Tables 2 and 3 show the mean bias, variance and the mean square error that we compute for each of the three estimators of $p_{a_0e_1}(s, s+4, 0, \Delta_v)$ and $p_{e_1d}(s, s+4, \Delta_u, \infty)$. Calculations are done with $K = 1,000$ replicated datasets. Similar results are found for $p_{a_0e_2}(s, s+4, 0, \Delta_v)$ and $p_{e_2d}(s, s+4, \Delta_u, \infty)$. They are detailed in the supplementary material. The simulations clearly indicate that the alternative estimators are more efficient than the first class of naive estimators introduced in Section 3.1, as they generally contain a significant bias even when the sample size increases. The specification used for C has an impact on the performance of estimators (3.10) and (3.13). This is emphasized by the results obtained with Scenario 2. Indeed, the bias of the estimated probability (3.10) for sample of size $n = 400$ remains at a low level and that of (3.10) is smaller than the alternative estimators in most of cases. These last results seem no longer verified asymptotically as the alternative estimators are still better for $p_{a_0e_1}(s, s+4, 0, \Delta_v)$ and are comparable for $p_{e_1d}(s, s+4, \Delta_u, \infty)$ with a sample of size $n = 1,500$ (additional simulation results not shown). On the other hand, the performances for alternative estimators are also slightly better in terms of mean square errors, while they are quite comparable regarding the variance indicator, except for some particular cases.

Both alternative estimators give very satisfactory results. The only real change is for small samples where both the variance and the mean square error indicators are better for $\widehat{p}^{*}_{e_1d}(s, s+4, \Delta_u, \infty)$, with an exception for $(s, t, \Delta_u) = (28.78, 32.78,]0, 2])$.

Additionally, we have also studied the performance of our estimators where C follows an independent Uniform distribution $\mathcal{U}[0, 90]$. The results are quite similar than those under Scenario 2, but with worst performance since censoring events are more frequent.

4.2 Comparison with a semi-Markov model

In insurance, the manner of observing the data is dependent on the claims management system which are designed on the terms of the contract. The states of the resulting multi-states model are based on the payment of the cash-flows. A bias can be induced when estimating this model if the risk of dying after the entry in disability which depends upon the duration time which can be unknown. The insurer observes instead the duration time in a contractual state as the health status of an insured comes to its knowledge when the claim is reported. The aim of this section consists in evaluating the bias which can appear if the estimated probabilities are based on the transition intensities of a semi-Markov model with a distorted duration time.

For that, we consider an irreversible illness-death model $\{X(t), t \geqslant 0\}$ where $X(t) \in \{a_0, e, d\}$

[5]Here, we scan the time on the basis of 2 time units. As estimation is done with relatively small-samples, we consider these time points to ensure that the transition probabilities are not too small and that the estimates are sufficient accuracy.

Table 2: Performance analysis for estimated transition probabilities from state a_0 to state e_1.

				$\widehat{p}_{a_0e_1}(s,t,0,\Delta_v)$			$\widehat{p}^{*}_{a_0e_1}(s,t,0,\Delta_v)$			$\breve{p}_{a_0e_1}(s,t,0,\Delta_v)$		
(s,t,Δ_v)	n	Censoring	$p_{a_0e_1}(s,t,0,\Delta_v)$	BIAS	VAR	MSE	BIAS	VAR	MSE	BIAS	VAR	MSE
(28.78,32.78,]0,2])	100	Scenario 1	0.045	12.77	1.20	1.36	-0.15	0.94	0.94	-0.10	0.94	0.94
	100	Scenario 2	0.045	1.42	1.09	1.10	-0.39	1.03	1.03	-0.40	1.03	1.03
	200	Scenario 1	0.045	14.19	0.50	0.70	0.36	0.42	0.42	0.36	0.42	0.42
	200	Scenario 2	0.045	1.60	0.62	0.62	0.30	0.52	0.52	0.32	0.52	0.52
	400	Scenario 1	0.045	11.95	0.34	0.49	-0.18	0.21	0.21	-0.15	0.21	0.21
	400	Scenario 2	0.045	0.94	0.32	0.32	0.30	0.28	0.28	0.30	0.28	0.28
(28.78,32.78,]2,4])	100	Scenario 1	0.023	9.12	0.55	0.64	-0.35	0.43	0.43	-0.37	0.43	0.43
	100	Scenario 2	0.023	0.50	0.66	0.66	-1.38	0.59	0.59	-1.32	0.60	0.60
	200	Scenario 1	0.023	9.61	0.24	0.33	-0.16	0.20	0.20	-0.16	0.20	0.20
	200	Scenario 2	0.023	0.72	0.34	0.34	-0.53	0.27	0.27	-0.52	0.27	0.27
	400	Scenario 1	0.023	8.50	0.20	0.27	-0.32	0.11	0.11	-0.31	0.11	0.11
	400	Scenario 2	0.023	0.21	0.17	0.17	-0.22	0.14	0.14	-0.22	0.14	0.14
(32.35,36.35,]0,2])	100	Scenario 1	0.069	32.96	3.12	4.20	0.35	3.02	3.02	0.25	3.02	3.02
	100	Scenario 2	0.069	5.06	3.28	3.30	0.96	2.82	2.82	1.05	2.84	2.84
	200	Scenario 1	0.069	30.34	1.88	2.80	-0.87	1.73	1.73	-0.80	1.70	1.70
	200	Scenario 2	0.069	1.97	1.67	1.67	-0.12	1.39	1.38	-0.16	1.39	1.39
	400	Scenario 1	0.069	29.42	1.12	1.98	0.25	0.83	0.83	0.31	0.83	0.83
	400	Scenario 2	0.069	2.11	0.75	0.76	0.36	0.63	0.63	0.34	0.63	0.63
(32.35,36.35,]2,4])	100	Scenario 1	0.052	25.80	3.18	3.84	0.34	2.02	2.02	0.92	2.02	2.02
	100	Scenario 2	0.052	5.85	2.35	2.38	2.12	2.00	2.00	2.39	2.04	2.05
	200	Scenario 1	0.052	27.35	1.55	2.30	0.51	1.11	1.11	0.45	1.10	1.10
	200	Scenario 2	0.052	3.95	1.22	1.23	2.27	0.99	1.00	2.30	1.00	1.00
	400	Scenario 1	0.052	26.33	1.12	1.82	2.32	0.57	0.57	2.37	0.56	0.57
	400	Scenario 2	0.052	2.43	0.73	0.74	0.64	0.56	0.56	0.59	0.56	0.56
(35.49,39.49,]0,2])	100	Scenario 1	0.087	64.76	7.09	11.27	0.86	15.33	15.32	6.09	13.97	13.99
	100	Scenario 2	0.087	10.40	10.61	10.71	-0.54	9.42	9.41	-0.26	9.53	9.52
	200	Scenario 1	0.087	57.30	6.17	9.45	-0.60	7.35	7.35	-0.62	7.28	7.27
	200	Scenario 2	0.087	9.56	5.22	5.31	-0.50	3.99	3.99	-0.44	4.02	4.01
	400	Scenario 1	0.087	57.32	3.20	6.48	-4.64	3.53	3.54	-4.34	3.51	3.53
	400	Scenario 2	0.087	2.20	2.92	2.92	-4.04	2.21	2.22	-4.01	2.21	2.22
(35.49,39.49,]2,4])	100	Scenario 1	0.099	76.76	9.20	15.08	-4.36	13.28	13.29	4.63	11.87	11.88
	100	Scenario 2	0.099	8.29	12.53	12.59	-3.91	10.81	10.81	-2.83	10.91	10.91
	200	Scenario 1	0.099	72.94	5.10	10.42	2.44	6.27	6.27	3.25	6.03	6.04
	200	Scenario 2	0.099	6.32	6.38	6.42	-1.24	4.99	4.99	-0.97	5.04	5.04
	400	Scenario 1	0.099	70.02	3.64	8.54	0.29	3.41	3.41	0.66	3.27	3.27
	400	Scenario 2	0.099	4.14	2.86	2.88	-1.58	2.18	2.18	-1.59	2.17	2.17

Note: This table contains the estimates bias (BIAS) $\times 10^3$, variance (VAR) $\times 10^3$ and mean square error (MSE) $\times 10^3$ with our non-parametric estimators. We compare the results at time $s = \tau_{.20}$, $s = \tau_{.40}$ and $s = \tau_{.60}$ for samples with size $n = 100$, $n = 200$ and $n = 400$. The sojourn time is comprised in $]0, 2]$ and in $]2, 4]$. The results are obtained with $K = 1,000$ Monte Carlo simulations.

Table 3: Performance analysis for estimated transition probabilities from state e_1 to state d.

				$\widehat{p}_{e_1d}(s,t,0,\Delta_v)$			$\widehat{p}^*_{e_1d}(s,t,0,\Delta_u)$			$\breve{p}_{e_1d}(s,t,0,\Delta_v)$		
(s,t,Δ_v)	n	Censoring	$p_{e_1d}(s,t,0,\Delta_v)$	BIAS	VAR	MSE	BIAS	VAR	MSE	BIAS	VAR	MSE
(28.78,32.78,]0,2])	100	Scenario 1	0.579	-209.92	121.94	165.88	106.33	168.70	179.83	-66.33	150.88	155.12
	100	Scenario 2	0.579	-36.38	176.12	177.27	143.10	187.82	208.11	-94.53	165.94	174.71
	200	Scenario 1	0.579	-198.86	96.80	136.25	24.18	93.46	93.95	0.33	94.55	94.45
	200	Scenario 2	0.579	-9.92	125.44	125.41	67.14	127.04	131.43	-5.49	123.26	123.16
	400	Scenario 1	0.579	-149.22	68.11	90.31	35.82	45.10	46.34	36.15	44.95	46.22
	400	Scenario 2	0.579	-8.58	65.47	65.48	30.64	57.59	58.47	23.30	57.96	58.44
(28.78,32.78,]2,4])	100	Scenario 1	0.369	-333.46	182.90	293.91	161.89	141.49	167.56	-312.03	196.56	293.73
	100	Scenario 2	0.369	-91.89	224.27	232.49	189.52	134.07	169.85	-380.66	177.29	322.01
	200	Scenario 1	0.369	-316.26	181.70	281.53	73.22	153.19	158.39	-173.71	196.46	226.44
	200	Scenario 2	0.369	-63.47	196.97	200.80	111.68	151.11	163.44	-214.52	205.20	251.01
	400	Scenario 1	0.369	-275.91	162.73	238.69	10.44	111.91	111.91	-55.82	130.70	133.68
	400	Scenario 2	0.369	-57.78	147.36	150.55	30.06	124.29	125.07	-85.83	151.49	158.70
(32.35,36.35,]0,2])	100	Scenario 1	0.526	-286.53	116.72	198.70	87.25	152.65	160.11	-34.65	160.61	161.65
	100	Scenario 2	0.526	-57.05	162.67	165.76	77.71	162.69	168.56	-70.62	161.03	165.86
	200	Scenario 1	0.526	-281.02	93.33	172.21	28.23	88.58	89.29	11.06	90.29	90.32
	200	Scenario 2	0.526	-40.23	100.70	102.22	16.61	94.40	94.58	-12.59	94.58	94.64
	400	Scenario 1	0.526	-243.19	66.71	125.78	7.78	41.74	41.76	3.52	42.66	42.63
	400	Scenario 2	0.526	-9.12	50.61	50.64	10.65	43.77	43.84	9.64	44.49	44.54
(32.35,36.35,]2,4])	100	Scenario 1	0.408	-338.95	173.34	288.06	178.86	143.72	175.56	-173.96	215.28	245.33
	100	Scenario 2	0.408	-9.06	201.81	201.69	183.22	140.34	173.77	-198.59	208.77	248.00
	200	Scenario 1	0.408	-317.80	162.41	263.25	99.46	125.36	135.12	-10.29	162.14	162.08
	200	Scenario 2	0.408	-1.19	153.55	153.40	106.41	126.48	137.67	-49.70	167.73	170.03
	400	Scenario 1	0.408	-288.69	135.59	218.80	38.62	84.50	85.91	39.10	84.34	85.78
	400	Scenario 2	0.408	9.60	95.86	95.85	47.11	81.41	83.55	22.87	88.07	88.50
(35.49,39.49,]0,2])	100	Scenario 1	0.447	-424.35	99.90	279.87	99.08	161.31	170.97	-122.18	187.57	202.31
	100	Scenario 2	0.447	-62.68	175.49	179.24	74.19	157.93	163.27	-109.99	178.18	190.10
	200	Scenario 1	0.447	-373.63	112.03	251.51	46.41	118.32	120.36	-2.80	127.82	127.70
	200	Scenario 2	0.447	-59.75	115.43	118.89	-6.48	104.89	104.83	-36.01	109.61	110.79
	400	Scenario 1	0.447	-365.66	85.79	219.42	-8.23	65.57	65.57	-11.07	64.72	64.78
	400	Scenario 2	0.447	-37.44	56.02	57.37	-8.59	47.23	47.26	-11.93	47.63	47.72
(35.49,39.49,]2,4])	100	Scenario 1	0.349	-452.29	144.20	348.62	156.81	125.40	149.86	-214.00	218.60	264.18
	100	Scenario 2	0.349	-62.07	188.42	192.08	108.88	137.45	149.17	-211.32	206.18	250.63
	200	Scenario 1	0.349	-470.81	128.23	349.76	77.92	122.39	128.34	-57.78	173.26	176.42
	200	Scenario 2	0.349	-31.91	133.58	134.46	50.38	108.82	111.25	-60.33	144.29	147.79
	400	Scenario 1	0.349	-435.69	124.08	313.78	16.98	95.03	95.23	14.25	94.41	94.52
	400	Scenario 2	0.349	-27.23	79.83	80.49	10.05	63.18	63.22	-1.70	67.73	67.66

Note: This table contains the estimates bias (BIAS) $\times 10^3$, variance (VAR) $\times 10^3$ and mean square error (MSE) $\times 10^3$ with our non-parametric estimators. We compare the results at time $s = \tau_{.20}$, $s = \tau_{.40}$ and $s = \tau_{.60}$ for samples with size $n = 100$, $n = 200$ and $n = 400$. The sojourn time is comprised in $]0, 2]$ and in $]2, 4]$. The results are obtained with $K = 1,000$ Monte Carlo simulations.

corresponds to the health status of an individual. State a_0 is the healthy state, state e_1 represents an illness or a group of illnesses which may cause the entry into a disability state, and state d represents the death state. When an insured is in state e, he has the possibility to report his claim (as defined by the insurance contract) after a certain time-lag. We assume that X is an homogeneous semi-Markov process where the failure times T_{a_0e}, T_{a_0d} and T_{ed} are independent and follow Weibull laws with parameters λ and ρ as defined in Table 4. The set-up for the right censoring process C, the number of simulations and the sample size is exactly as in Section 4.1.

Table 4: Simulation parameters used for the semi-Markov specification.

Parameter	T_{a_0e}	T_{a_0d}	T_{ed}
Scale λ	35	50	5
Shape ρ	8	3	0.5

Note: This table displays the simulation parameters used with Weibull distributions Wei (λ, ρ).

When an individual enters in state e, his health status is diagnosed and reported to the insurer after a random lag-time. It is unobserved until $T_{ee'}$ where e' corresponds to the disability state triggering the payment of benefits in accordance with the contract. This latent time is simulated using a independent uniform distribution $\mathcal{U}[T_{a_0e}, T_{ed}]$. From the insurer's point of view, the quantities of interest are the transition probability from a_0 to e' and not to e. With no loss of generality, we consider the probabilities $p_{a_0e'}(s, s+2) = p_{a_0e'}(s, s+2, 0, [0, t-s]) = \mathbb{P}(X_{s+2} = e' \mid X_s = a_0)$ for $s \geqslant 0$. Without right censoring, these quantities can be easily estimated not parametrically. Conversely, our estimators provide direct estimates of these quantities.

In this framework, let us consider now a second homogeneous semi-Markov model $\{X'(t), t \geqslant 0\}$ where $X'(t) \in \{a_0, e', d\}$ for estimating the probabilities of interest on observable variables. The duration laws for each transition are assumed to be Weibull Wei (ν, σ) where (ν, σ) are two parameters to estimate. Since the data are initially simulated using Weibull distributions and this law is quite flexible, one would expect a good quality in terms of goodness of fit. In particular, this approach fits well without the introduction of a latent period before the claim notification. Thus, we compute the transition probabilities by estimating first the transition intensities of the semi-Markov model and then by evaluating the following integral using numerical integration with a trapezoid rule and 1,000 samples

$$\mathbb{P}\left(X_{s+2} = e' \mid X_s = a_0\right) = \int_s^{s+2} \bar{p}_{a_0a_0}(s, \tau, 0)\, \mu_{a_0e'}(\tau)\, \bar{p}_{e'e'}(\tau, t, 0)\, d\tau.$$

The results of this comparison is presented in Table 5. Our estimators $\widehat{p}_{a_0e'}(s, s+2)$ and $\breve{p}_{a_0e'}(s, s+2)$ outperform the semi-Markov estimators in terms of bias. Not surprisingly, the semi-Markov estimators do show a smaller variance in many cases since the model is parametric. The superiority of $\widehat{p}_{a_0e'}(s, s+2)$ compared to semi-Markov estimators is less clear as this estimator can be seriously biased as noted in Section 4.1.

5 Application to LTC insurance data

This Section describes our LTC insurance dataset and discusses the results obtained with our non-parametric estimators for transition probabilities. Here, we compare these estimators and measure

Table 5: Comparison with the estimated transition probabilities from state a_0 to state e with a homogeneous semi-Markov model.

				$\widehat{p}_{a_0e'}(s,s+2)$			$\widehat{p}_{a_0e'}(s,s+2)$			$\breve{p}_{a_0e'}(s,s+2)$			Semi-Markov model		
(s,t)	n	Censoring	$p_{a_0e'}(s,s+2)$	BIAS	VAR	MSE	BIAS	VAR	MSE	BIAS	VAR	MSE	BIAS	VAR	MSE
(31.30,33.30)	100	Scenario 1	0.059	7.06	1.80	1.85	1.03	1.07	1.07	0.99	1.07	1.07	-19.55	0.42	0.80
(31.30,33.30)	100	Scenario 2	0.059	-1.42	1.57	1.57	-1.76	1.46	1.47	-1.79	1.46	1.46	-3.57	0.30	0.31
(31.30,33.30)	200	Scenario 1	0.059	6.70	0.88	0.93	0.22	0.53	0.53	0.21	0.53	0.53	-20.99	0.27	0.71
(31.30,33.30)	200	Scenario 2	0.059	0.72	0.78	0.78	0.70	0.70	0.70	0.72	0.70	0.70	-0.38	0.14	0.14
(31.30,33.30)	400	Scenario 1	0.059	6.25	0.51	0.55	0.12	0.27	0.27	0.13	0.27	0.27	-19.09	0.19	0.56
(31.30,33.30)	400	Scenario 2	0.059	-0.54	0.35	0.35	-0.44	0.33	0.33	-0.44	0.33	0.33	1.30	0.06	0.06
(35.16,37.16)	100	Scenario 1	0.112	25.79	6.20	6.86	-0.52	4.94	4.93	-0.48	4.88	4.87	-24.20	2.10	2.69
(35.16,37.16)	100	Scenario 2	0.112	-0.74	4.60	4.59	-0.92	4.23	4.23	-0.84	4.25	4.25	17.29	1.35	1.65
(35.16,37.16)	200	Scenario 1	0.112	20.85	4.19	4.62	-1.36	2.22	2.22	-1.27	2.21	2.21	-28.41	1.46	2.26
(35.16,37.16)	200	Scenario 2	0.112	0.34	2.31	2.30	0.02	2.13	2.12	0.07	2.12	2.12	27.32	0.56	1.30
(35.16,37.16)	400	Scenario 1	0.112	21.82	2.35	2.82	-1.16	1.13	1.13	-1.09	1.11	1.11	-23.18	1.24	1.78
(35.16,37.16)	400	Scenario 2	0.112	-0.77	1.16	1.16	-0.93	1.02	1.02	-0.96	1.02	1.02	32.38	0.22	1.27
(38.90,40.90)	100	Scenario 1	0.156	71.84	25.78	30.92	-4.21	24.34	24.33	$*$	$*$	$*$	-13.23	10.48	10.65
(38.90,40.90)	100	Scenario 2	0.156	-0.14	13.16	13.15	-2.15	11.70	11.70	-1.74	11.76	11.75	24.63	3.95	4.56
(38.90,40.90)	200	Scenario 1	0.156	68.03	16.13	20.74	2.59	10.36	10.36	$*$	$*$	$*$	-30.04	8.37	9.27
(38.90,40.90)	200	Scenario 2	0.156	-2.54	6.42	6.42	-3.79	6.06	6.07	-3.62	6.03	6.04	43.63	1.55	3.46
(38.90,40.90)	400	Scenario 1	0.156	62.88	9.42	13.37	-6.48	5.23	5.27	$*$	$*$	$*$	-20.33	8.34	8.74
(38.90,40.90)	400	Scenario 2	0.156	-4.90	3.30	3.32	-4.97	2.93	2.95	-4.99	2.92	2.94	53.11	0.56	3.38

Note: This table contains the estimates bias (BIAS) $\times 10^3$, variance (VAR) $\times 10^3$ and mean square error (MSE) $\times 10^3$ with our non-parametric estimators. We compare the results at time $s = \tau_{.20}$, $s = \tau_{.40}$ and $s = \tau_{.60}$ for samples with size $n = 100$, $n = 200$ and $n = 400$. The results are obtained with $K = 1,000$ Monte Carlo simulations. $*$ indicates that there is a lack of data for a particular time interval and $\breve{p}_{a_0e'}(s,s+2)$ can not be estimated.

their uncertainty with a non-parametric bootstrap procedure.

5.1 Data description

The dataset that we analyze here is drawn from the database of a French LTC insurer. The corresponding policy offers fixed benefits for elderly people and covers the insured's residual lifetime. The disability states only include the most severe degrees of disability and the contract does not allow to recover from a disability state. The insured pays his premium as long as he is in the healthy state. The guarantee is terminated or reduced after a fixed period of time when the policyholder stops paying the premium. Benefits may also be paid in case of death before entry in dependency or when the contract is reduced.

This dataset is almost the same as that used to estimate the probabilities of entry into dependency in Guibert and Planchet (2014). It is also studied by Tomas and Planchet (2013) with an adaptive non-parametric approach for smoothing the survival law of LTC claimants, but without distinguishing the effect of each disease causing the entry into dependency. In this last study, the authors compute the monthly death rates for the population of LTC heavy claimants and exhibit significant differences for duration times, indicating that the lifetime after entry into dependency clearly depends upon the age and the time elapsed since the entry. In addition, we have fitted a Cox semi-Markov model as a preliminary test to assess the Markov assumption, which returns the duration time as a significant factor. This means that the Markov assumption is not satisfied. Comparing to these previous studies, we have improved the data by adding the individual cause-specific living path after entry into dependency.

The data are longitudinal with independent right-censoring and left-truncation. Most of the time, censoring occurs at the end of the observation period, but other administrative causes can

exist. For example, a policyholder may be censored if he refuses to continue paying his premiums and loses his rights. Left-truncation is induced by the entry date of a policyholder into the insurance portfolio. As the terms of the contract provide a lifetime coverage period, there is apparently no link between the truncation and the censoring times. For these reasons, the independence assumption for left-truncation and right censoring is considered to be satisfied.

The period of observation stretches from the beginning of 1998 to the end of 2010, and the range of ages is 65-90. Note that the definition of the disability states is relatively uniform over the studied period. We observe around 210,000 contracts and 68% of insured are censored before entering into dependency. The data are split into 4 groups of pathologies: neurological pathologies e_1 (censoring rates=40%) including strokes, various pathologies e_2 (censoring rates=36%), terminal cancers e_3 (censoring rates=6%) and dementia e_4 (censoring rates=45%) including Alzheimer's disease. About 12,800 entries in dependency are observed. Considering death d_1 and reduction d_2 as the other exit causes from the initial state, the multi-state structure of our dataset is shown in Figure 4. The state d_1 can be reached from all the disability states. We do not consider any covariate in this application.

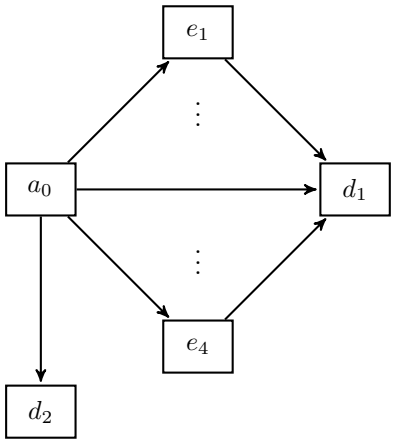


Figure 4: Multi-state structure of LTC data.

5.2 Estimation results for transition probabilities

Now, we perform the estimation of transition probabilities with each method presented in Section 3, i.e. for $e \in \{e_1, \ldots, e_4\}$:

- Method 1: $\widehat{p}_{a_0 e}(s,t,0,\Delta_v)$ and $\widehat{p}_{ed_1}(s,t,\Delta_u,\infty)$,
- Method 2: $\widehat{p}^{*}_{a_0 e}(s,t,0,\Delta_v)$ and $\widehat{p}^{*}_{ed_1}(s,t,\Delta_u,\infty)$,
- Method 3: $\breve{p}^{*}_{a_0 e}(s,t,0,\Delta_v)$ and $\breve{p}^{*}_{ed_1}(s,t,\Delta_u,\infty)$.

For brevity's sake, we do not show the other transition probabilities of interest described in Section 2, as studying (3.3) and (3.6) is sufficient to illustrate the methodological issue raised in this paper.

Figure 5 displays the estimates of the annual probabilities $p_{a_0 e}(s, s+1, 0, \Delta_v)$ of becoming dependent for an insured between the ages of s and $s+1$, and staying one month ($\Delta_v =]0, \frac{1}{12}]$) in a disability state[6]. Figure 6 shows similarly the entry probabilities with a sojourn time between five and six months, along the same range of ages. For each disease, transition probabilities globally

[6] A monthly time step is chosen for discretization.

grow over time, with an acceleration after age 80. We note that the alternative estimators (Method 2 and Method 3) are virtually identical and are more robust than those calculated with Method 1. In most cases, the estimates related to Method 1 are smaller than its competitors, expect for some pics after the age of 80, suggesting than this method generates a negative bias here.

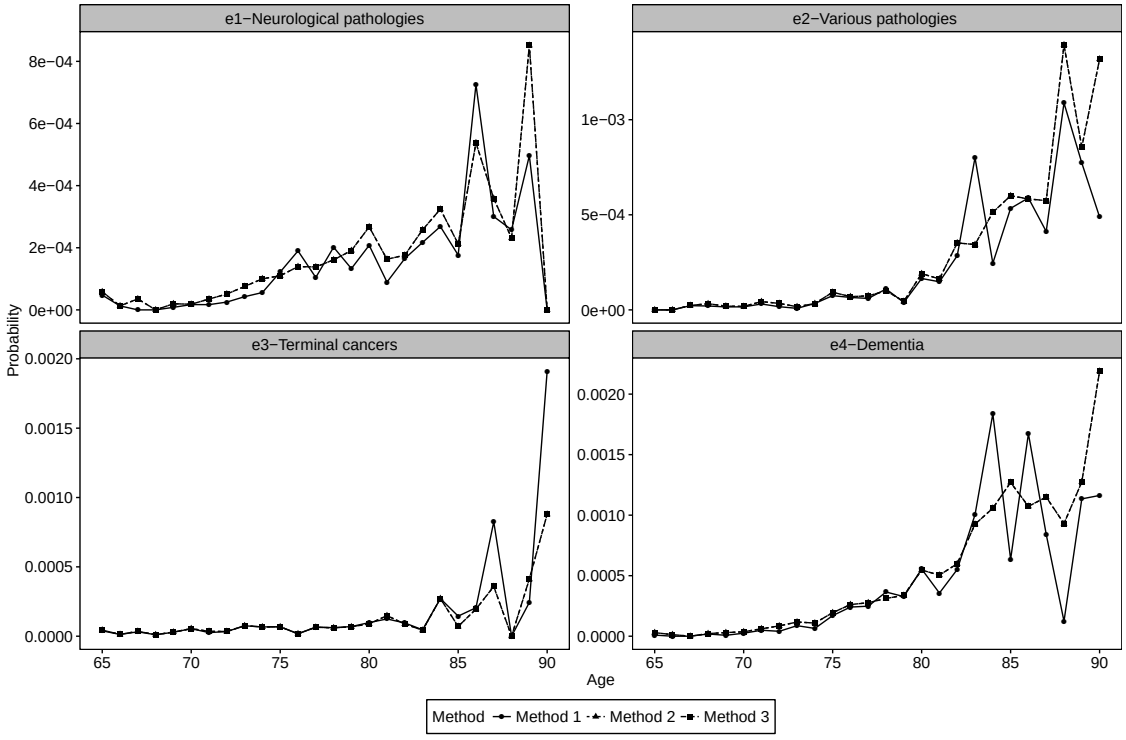


Figure 5: Annually estimated transition probabilities from the healthy state a_0 to the dependency states $e_1, \ldots, e_4$ with Method 1, Method 2 and Method 3. The range of ages is 65-90 and the sojourn time of one month.

Then, we analyze the estimated death probabilities from each disability state $p_{ed_1}(s, s+1/12, \Delta_u, \infty)$, $e = e_1, \ldots, e_4$, on a monthly basis for each method. Monthly rates are computed since practitioners are familiar with these quantities. As data are sparse for some time points, point estimates of transition probabilities may be erratic. For this reason, we choose to report a simple integrated version of these probabilities as follows

$$p_{ed_1}(\lfloor s \rfloor, \lfloor s \rfloor + 1/12, \Delta_u, \infty) = \sum_{k=0}^{1/\lambda - 1} \lambda p_{ed_1}(s + k\lambda, s + 1/12 + k\lambda, \Delta_u, \infty),$$

for fixed values of s and Δ_u, and with a monthly time step $\lambda = 1/12$. This provides an average of the monthly death probabilities for individuals entering in dependency with the same (integer) age. Note that a suitable choice for the weights and the calculation time points could be found, but we do not develop this point further to keep the comparison of methods as simple as possible. Figure 7 and Figure 8 present these integrated transition probabilities for duration u varying between one month and six months, and with ages of occurrence equal to 75 and 80 respectively. We limit our analysis to these windows as censoring becomes important beyond a twelve months duration period for causes e_1, e_2 and e_4, and observations are scarce for terminal cancers exceeding a period of six months. We find that the residual lifetime (after an individual enters into dependency) is very different from one type of disease to another. In particular, the model shows extreme death

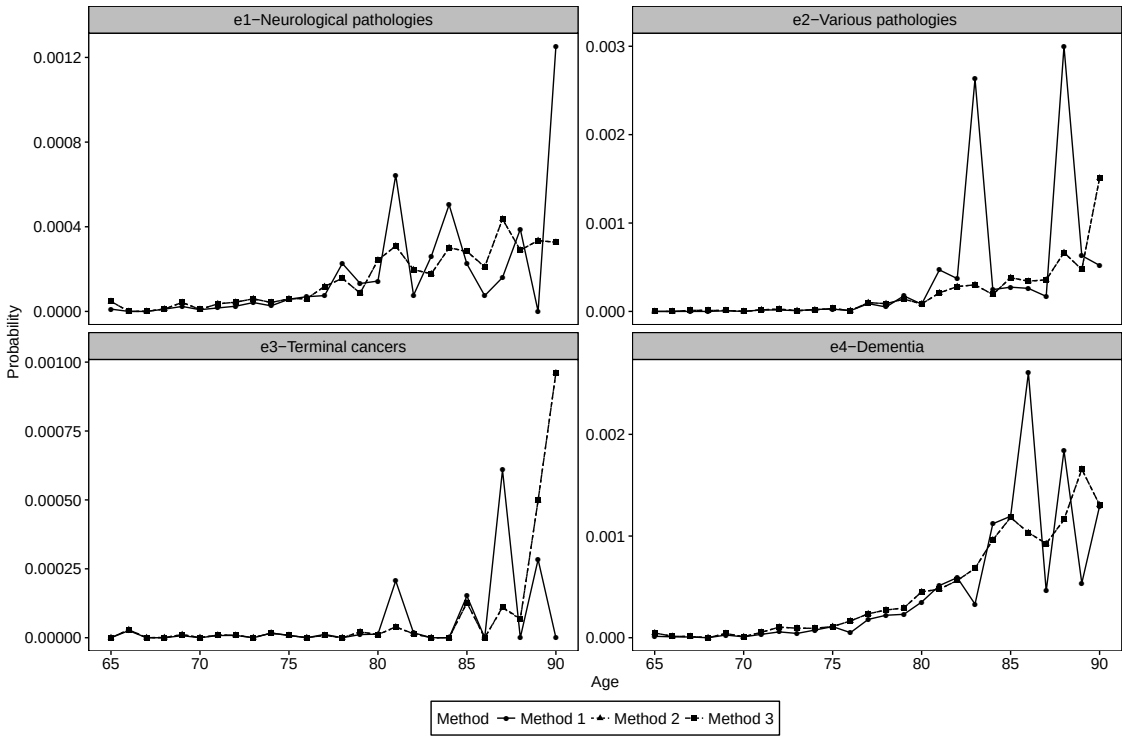


Figure 6: Annually estimated transition probabilities from the healthy state a_0 to the dependency states $e_1, \ldots, e_4$ with Method 1, Method 2 and Method 3. The range of ages is 65-90 and the sojourn time between five and six months.

probabilities for terminal cancers for the first six months. The comparison between estimation methods provides interesting results. For causes e_1, e_2 and e_4, both alternative estimators are quite similar, with an exception for e_2 at age 75 and for $u = 2$. For terminal cancers, results given by Method 3 seem to be underestimated, when the age of occurrence is 75. Conversely, results with Method 1 and Method 2 are close. Certainly, this is explained by a low censoring rate and a relatively small number of policyholders at this age. Additionally, estimates are close to each other at age 80, when increasing the exposure. This finding is comparable to results described in Section 4 with Scenario 2.

We now report some point estimates for $p_{a_0e}(s, s, +1, 0, \Delta_v)$ in Table 6 and $p_{ed_1}(\lfloor s \rfloor, \lfloor s \rfloor + 1/12, \Delta_u, \infty)$ in Table 7 along with 95% confidence intervals at age $s = 75$ and $s = 80$. For that, the asymptotic variance is obtained with 500 bootstrap samples and the confidence intervals are deduced using the normal approximation. For both probabilities, the sojourn time in disability states are $\Delta_v = 1/12$ and $\Delta_v = 1/2$, and $\Delta_u = 1/12$ and $\Delta_u = 1/2$. We compare the results obtained with Method 1, Method 2 and Method 3. In Table 6, results obtained with alternative estimators (Method 2 and Method 3) are virtually the same and are more robust than those of Method 1 in most cases, except for cause e_2 at age 75. Estimates of $p_{ed_1}(\lfloor s \rfloor, \lfloor s \rfloor + 1/12, \Delta_u, \infty)$ are highly uncertain as the amount of available data is insufficient. While remaining wary and critical due to this lack of reliability, we observe however that the difference between the two alternative estimators remains small, except for cause e_3 for a duration of six months where data quality is probably poor. Ignoring results for cause e_3, note again that the 95% confidence intervals are generally larger for estimator $\widehat{p}_{ed_1}(\lfloor s \rfloor, \lfloor s \rfloor + 1/12, \Delta_u, \infty)$, especially at age 80.

Table 6: Annually estimated probabilities from the healthy state with confidence intervals.

(s, Δ_v)	Method 1	Method 2	Method 3
	e1-Neurological pathologies		
(75,1/12)	0.0123	0.011	0.011
	$(0.0039, 0.0207)$	$(0.0049, 0.0171)$	$(0.0049, 0.0171)$
(75,1/2)	0.0057	0.0059	0.0059
	$(0.0003, 0.0111)$	$(0.0013, 0.0104)$	$(0.0013, 0.0105)$
(80,1/12)	0.0207	0.0266	0.0266
	$(0.0082, 0.0332)$	$(0.0152, 0.0381)$	$(0.0152, 0.0381)$
(80,1/2)	0.0142	0.0241	0.0241
	$(0.0017, 0.0267)$	$(0.0138, 0.0344)$	$(0.0138, 0.0345)$
	e2-Various pathologies		
(75,1/12)	0.0074	0.0091	0.0091
	$(0.0023, 0.0125)$	$(0.0036, 0.0147)$	$(0.0036, 0.0147)$
(75,1/2)	0.0026	0.0034	0.0034
	$(0, 0.0055)$	$(0, 0.0068)$	$(0, 0.0068)$
(80,1/12)	0.0166	0.019	0.019
	$(0.0052, 0.028)$	$(0.0094, 0.0286)$	$(0.0094, 0.0286)$
(80,1/2)	0.0084	0.0085	0.0086
	$(0.0014, 0.0154)$	$(0.0024, 0.0147)$	$(0.0024, 0.0147)$
	e3-Terminal cancers		
(75,1/12)	0.007	0.0069	0.0069
	$(0.0023, 0.0117)$	$(0.0022, 0.0115)$	$(0.0022, 0.0115)$
(75,1/2)	0.0008	0.0008	0.0008
	$(0, 0.0026)$	$(0, 0.0025)$	$(0, 0.0026)$
(80,1/12)	0.0098	0.0089	0.0089
	$(0.0026, 0.017)$	$(0.0025, 0.0153)$	$(0.0025, 0.0153)$
(80,1/2)	0.0013	0.0012	0.0012
	$(0, 0.0039)$	$(0, 0.0035)$	$(0, 0.0036)$
	e4-Dementia		
(75,1/12)	0.0169	0.0192	0.0192
	$(0.009, 0.0248)$	$(0.0112, 0.0272)$	$(0.0112, 0.0272)$
(75,1/2)	0.0113	0.0112	0.0112
	$(0.0041, 0.0184)$	$(0.0053, 0.0172)$	$(0.0053, 0.0172)$
(80,1/12)	0.0546	0.0543	0.0543
	$(0.0279, 0.0814)$	$(0.0378, 0.0708)$	$(0.0378, 0.0708)$
(80,1/2)	0.0351	0.045	0.0451
	$(0.0117, 0.0586)$	$(0.0295, 0.0606)$	$(0.0295, 0.0606)$

This table gives the estimated transition probabilities $p_{a_0 e}(s, s, +1, 0, \Delta_v) \times 10^2$ for each dependency state $e_1, \ldots, e_4$ with the 95% confidence interval in parentheses computed with 500 bootstrap replications. The results are calculated at ages $s = 75$ and $s = 80$ and with durations $\Delta_v = 1$ month and $\Delta_v = 6$ months.

Table 7: Estimates of integrated monthly death probabilities with confidence intervals.

(s, Δ_v)	Method 1	Method 2	Method 3
	e1-Neurological pathologies		
(75,1)	3.116	2.5743	2.4488
	$(0, 7.7344)$	$(0, 6.4165)$	$(0, 5.9942)$
(75,6)	5.4247	4.2354	4.1023
	$(0, 10.9402)$	$(0, 8.713)$	$(0, 8.2716)$
(80,1)	1.5662	1.4368	1.4174
	$(0, 3.9331)$	$(0, 3.4276)$	$(0, 3.3809)$
(80,6)	4.9534	4.1229	4.0652
	$(0.6793, 9.2275)$	$(0.9018, 7.3441)$	$(0.8695, 7.2608)$
	e2-Various pathologies		
(75,1)	7.2656	5.0384	4.6507
	$(0, 15.3139)$	$(0, 10.759)$	$(0, 9.7527)$
(75,6)	1.7813	1.3181	1.3002
	$(0, 5.1237)$	$(0, 3.7893)$	$(0, 3.6938)$
(80,1)	6.0748	4.7783	4.7729
	$(1.6197, 10.5298)$	$(1.4377, 8.1189)$	$(1.4698, 8.076)$
(80,6)	1.9048	1.391	1.3694
	$(0, 4.9755)$	$(0, 3.3912)$	$(0, 3.3425)$
	e3-Terminal cancers		
(75,1)	18.501	17.5142	9.5096
	$(9.8504, 27.1516)$	$(9.32, 25.7084)$	$(3.378, 15.6411)$
(75,6)	9.3892	7.5959	3.5219
	$(0, 19.3642)$	$(0, 16.5577)$	$(0, 9.9758)$
(80,1)	30.6836	31.1928	29.9673
	$(21.1441, 40.2231)$	$(21.5775, 40.808)$	$(20.5805, 39.3541)$
(80,6)	15.8069	15.7636	3.1259
	$(2.2186, 29.3952)$	$(2.4011, 29.126)$	$(0, 8.2889)$
	e4-Dementia		
(75,1)	0.8661	0.7018	0.6986
	$(0, 2.0728)$	$(0, 1.6663)$	$(0, 1.6587)$
(75,6)	1.0705	0.9589	0.9856
	$(0, 2.6902)$	$(0, 2.375)$	$(0, 2.444)$
(80,1)	1.0176	0.9759	0.9606
	$(0, 2.1147)$	$(0, 1.9765)$	$(0, 1.9455)$
(80,6)	2.644	2.2492	2.2358
	$(0.6634, 4.6246)$	$(0.7096, 3.7887)$	$(0.702, 3.7697)$

This table gives the estimated transition probabilities $p_{ed_1}(\lfloor s \rfloor, \lfloor s \rfloor + 1/12, \Delta_u, \infty) \times 10^2$ for each dependency state $e_1, \ldots, e_4$ with the 95% confidence interval in parentheses computed with 500 bootstrap replications. The results are calculated at ages $s = 75$ and $s = 80$ and with durations $\Delta_v = 1$ month and $\Delta_v = 6$ months.

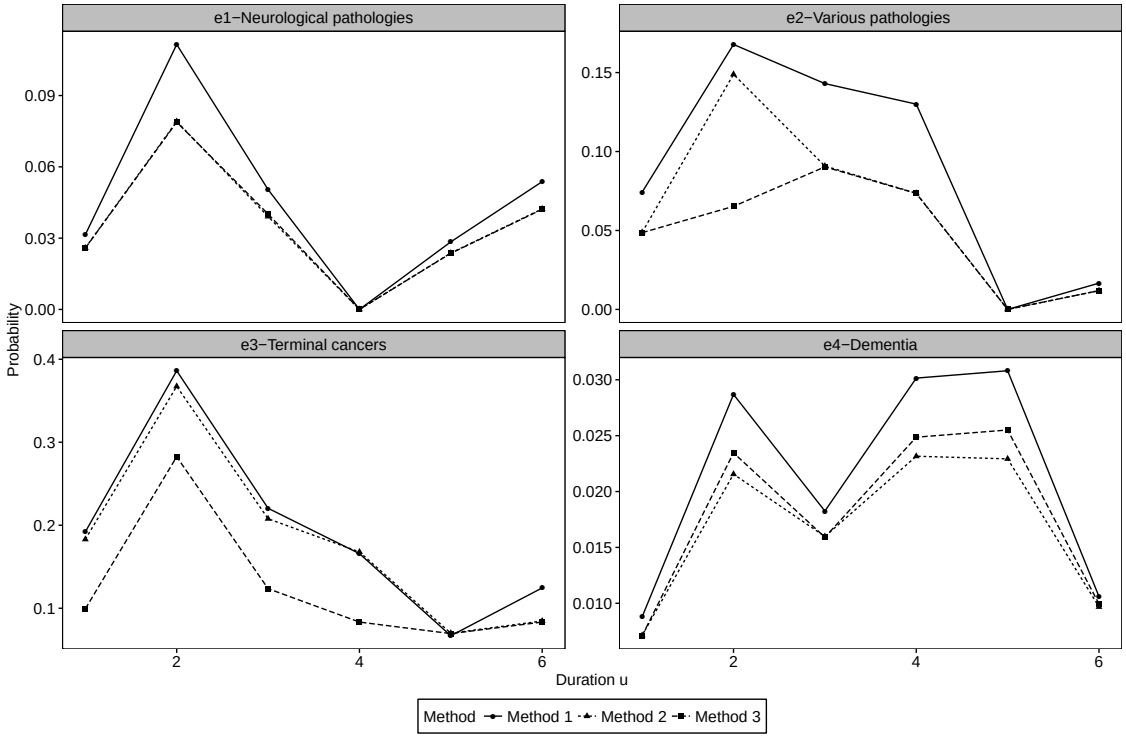


Figure 7: Estimates of integrated monthly death probabilities corresponding to entry in dependency at age 75 for each disability state $e_1, \ldots, e_4$. Estimation are done with Method 1, Method 2 and Method 3. Duration intervals are $\Delta_u =]0, 1/12],]1/12, 1/6], \ldots,]5/12, 1/2]$.

6 Discussion

This paper focuses on the non-parametric analysis of particular acyclic multi-state models relaxing the Markov assumption. Building on a competing risks set-up distinguishing two different lifetimes in presence of independent right-censoring, we propose estimators for some specific targets corresponding to probabilities of paying cash-flows arising from a LTC insurance contract. Our approach can also be extended to other health insurance contracts, like disability guarantees, with a similar multi-state structure when data are observed in continuous time. Such estimators can be used to check any assumption usually made in applications with observed data. A first set of estimators can be seen as Aalen-Johansen integrals for competing risks data using special indicator functions. These estimators can be affected by a bias problem due to censoring that we correct by deriving two alternatives, built following the de Uña-Álvarez and Meira-Machado (2015)' methodology. The asymptotic properties can be derived by adapting the classical results obtained for Kaplan-Meier integrals to competing risks data, and we show how estimate the model with additional left-truncation.

The simulations demonstrate the relevance of our approach to estimate particular transition probabilities, and suggest better performances for alternative estimators in terms of both bias and variance when the lifetime maximum value is censored. As an illustration, we compared our estimators adapted to actuarial needs on LTC insurance data, subjected to left-truncation and right censorship where the Markov assumption is not verified. In this situation, the both alternative are preferred to the initial set of estimators as they are more robust, in particular for higher ages exposed to heavily censoring. Variances and asymptotic confidence intervals are also estimated applying a simple non-parametric bootstrap technique. Some of our transition probabilities can

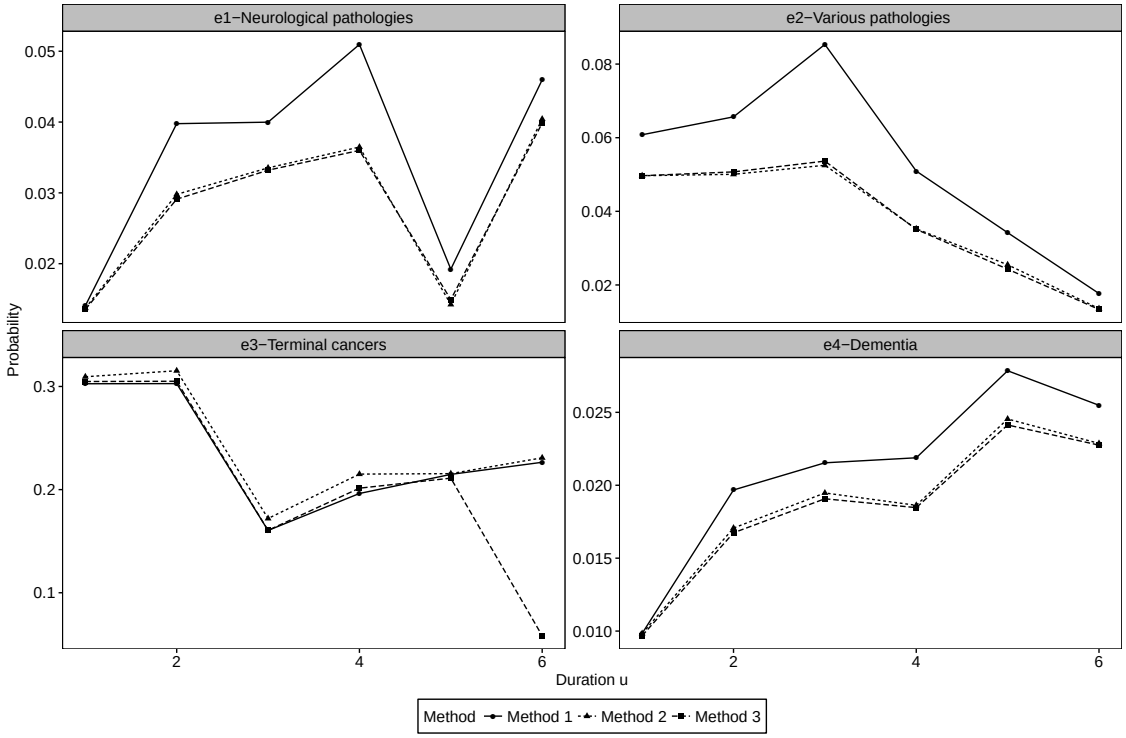


Figure 8: Estimates of integrated monthly death probabilities corresponding to entry in dependency at age 80 for each disability state $e_1, \ldots, e_4$. Estimations are done with Method 1, Method 2 and Method 3. Duration intervals are $\Delta_u =]0, 1/12],]1/12, 1/6], \ldots,]5/12, 1/2]$.

be robustly estimated with a semi-Markov model. However, we explain how the semi-Markov assumption can be violated if the duration time in an intermediate state is partially observed, e.g. when the time between the deterioration of the insured's health status and the claim notification is unknown and sufficiently important. In such a case, we show that a significant bias may appear when calculating the transition probabilities based on the estimates of a semi-Markov model. By contrast, our alternative estimators remain robust and accurate as they do not depend on the Markov assumption.

Our non-parametric estimators can also be applicable with observed discrete covariates. However, it may be difficult to apply it directly with lots of covariates due to the dimension of the problem. An outlook for future investigations consist in analysis non-parametric and semi-parametric regressions with the framework introduced in this paper. In particular, direct regression approaches developed by Andersen et al. (2003), and used for e.g. by Helms et al. (2005), could be considered as an interesting alternative to the classical Poisson regression approaches for transition intensities. Finally, there is another issue in constructing some goodness-of-fit tests which requires to compute appropriate critical values. We have considered a classic Efron's bootstrap technique (Efron, 1981) with resampling approach for our applications. Nevertheless with survival data in presence of censoring, this methodology is generally chosen since it is more consistent. These subjects are left for future works.

Acknowledgements

We are grateful to Olivier Lopez, Christian Robert and anonymous referees for the valuable advice and suggestions which they provided in relation to the first draft of this paper. The authors are also very thankful to Pierre-Emmanuel Darpeix for his helpful comments. This work benefited from the financial support of the ANR project "LoLitA" (ANR-13-BS01-0011). A substantial portion of this work was done when the first author was affiliated to the Autorité de Contrôle Prudentiel et de Résolution. The opinions expressed do not necessarily reflect views of the Autorité de Contrôle Prudentiel et de Résolution.

Supplementary material

Additional information and proofs referenced in Section 3 are available in supplementary material attached to this manuscript. This material also comprises additional simulation results related to Section 4.

References

Aalen, O. O. and Johansen, S. (1978). An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scandinavian Journal of Statistics* 5.3, pp. 141–150.

Allignol, A., Beyersmann, J., Gerds, T., and Latouche, A. (2014). A competing risks approach for nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis*, pp. 1–19. DOI: 10.1007/s10985-013-9269-1.

Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research* 11.2, pp. 91–115. DOI: 10.1191/0962280202SM276ra.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). Statistical Models Based on Counting Processes. Springer Series in Statistics. Springer-Verlag New York Inc.

Andersen, P. K., Klein, J. P., and Rosthøj, S. (2003). Generalised linear models for correlated pseudo-observations with applications to multi-state models. *Biometrika* 90.1. 00128, pp. 15–27. DOI: 10.1093/biomet/90.1.15.

Azarang, L., de Uña-Álvarez, J., and Stute, W. (2015). The Jackknife estimate of covariance of two Kaplan–Meier integrals with covariables. *Statistics* 49.5, pp. 1005–1025. DOI: 10.1080/02331888.2014.960871.

Beyersmann, J., Schumacher, M., and Allignol, A. (2011). Competing Risks and Multistate Models with R. Use R ! Springer.

Beyersmann, J., Termini, S. D., and Pauly, M. (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, pp. 387–402. DOI: 10.1111/j.1467-9469.2012.00817.x.

Biessy, G. (2015). Long-Term Care Insurance: a Multi-State Semi-Markov Model to Describe the Dependency Process for Elderly People. *Bulletin Français d'Actuariat* 15.29, pp. 41–74.

Buchardt, K., Møller, T., and Schmidt, K. B. (2014). Cash flows and policyholder behaviour in the semi-Markov life insurance setup. *Scandinavian Actuarial Journal*, pp. 1–29. DOI: 10.1080/03461238.2013.879919.

Cheng, Y., Fine, J. P., and Kosorok, M. R. (2007). Nonparametric Association Analysis of Bivariate Competing-Risks Data. *Journal of the American Statistical Association* 102.480, pp. 1407–1415. DOI: 10.1198/016214507000001157.

Christiansen, M. C. (2012). Multistate models in health insurance. *Advances in Statistical Analysis* 96.2, pp. 155–186. DOI: 10.1007/s10182-012-0189-2.

Courbage, C. and Roudaut, N. (2011). Long-term care insurance: The French example. *European Geriatric Medicine* 2.1, pp. 22–25. DOI: 10.1016/j.eurger.2010.11.006.

Currie, I. D., Durban, M., and Eilers, P. H. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling* 4.4. 00308, pp. 279–298. DOI: 10.1191/1471082X04st080oa.

Czado, C. and Rudolph, F. (2002). Application of survival analysis methods to long-term care insurance. *Insurance: Mathematics and Economics* 31.3, pp. 395–413. DOI: 10.1016/S0167-6687(02)00186-5.

De Uña-Álvarez, J. and Meira-Machado, L. (2015). Nonparametric estimation of transition probabilities in the non-Markov illness-death model: A comparative study. *Biometrics*. DOI: 10.1111/biom.12288.

De Wreede, L. C., Fiocco, M., and Putter, H. (2011). mstate: An R Package for the Analysis of Competing Risks and Multi-State Models. *Journal of Statistical Software* 38.7. DOI: 10.18637/jss.v038.i07.

Deléglise, M. P., Hess, C., and Nouet, S. (2009). Tarification, provisionnement et pilotage d'un portefeuille Dépendance. *Bulletin Français d'Actuariat* 9.17, pp. 70–108.

Denuit, M. and Robert, C. (2007). Actuariat des assurances de Personnes - Modélisation, tarification et provisionnement. Assurance Audit Actuariat. Paris: Economica.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* 7.1, pp. 1–26. DOI: 10.1214/aos/1176344552.

Efron, B. (1981). Censored Data and the Bootstrap. *Journal of the American Statistical Association* 76.374, pp. 312–319. DOI: 10.1080/01621459.1981.10477650.

Fong, J. H., Shao, A. W., and Sherris, M. (2015). Multistate Actuarial Models of Functional Disability. *North American Actuarial Journal* 19.1, pp. 41–59. DOI: 10.1080/10920277.2014.978025.

Fuino, M. and Wagner, J. (2017). Long-term care models and dependence probability tables by acuity level: new empirical evidence from Switzerland. Working Paper. University of Lausanne.

Gauzère, F., Commenges, D., Barberger-Gateau, P., Letenneur, L., and Dartigues, J.-F. (1999). Maladie et dépendance - Description des évolutions par des modèles multi-états. *Population* 54.2, pp. 205–222. DOI: 10.2307/1534898.

Guibert, Q. and Planchet, F. (2014). Construction de lois d'expérience en présence d'évènements concurrents – Application à l'estimation des lois d'incidence d'un contrat dépendance. *Bulletin Français d'Actuariat* 13.27, pp. 5–28.

Haberman, S. and Pitacco, E. (1998). Actuarial Models for Disability Insurance. 1st ed. Chapman and Hall/CRC.

Helms, F., Czado, C., and Gschlößl, S. (2005). Calculation of LTC Premiums Based on Direct Estimates of Transition Probabilities. *ASTIN Bulletin: The Journal of the International Actuarial Association* 35.02, pp. 455–469. DOI: 10.1017/S0515036100014331.

Hoem, J. M. (1972). Inhomogeneous semi-Markov processes, select actuarial tables, and duration-dependence in demography. *Population Dynamics*, pp. 251–296. DOI: 10.1016/B978-1-4832-2868-6.50013-8.

Hougaard, P. (1999). Multi-state Models: A Review. *Lifetime Data Analysis* 5.3, pp. 239–264. DOI: 10.1023/A:1009672031531.

— (2001). Analysis of Multivariate Survival Data. Statistics for Biology and Health. Springer-Verlag New York Inc.

Janssen, J. and Manca, R. (2007). Semi-Markov Risk Models for Finance, Insurance and Reliability. Springer-Verlag New York Inc.

Król, A. and Saint-Pierre, P. (2015). SemiMarkov: An R Package for Parametric Estimation in Multi-State Semi-Markov Models. *Journal of Statistical Software* 66.6, pp. 1–16.

Lando, D. and Skødeberg, T. M. (2002). Analyzing rating transitions and rating drift with continuous observations. *Journal of Banking & Finance* 26.2–3, pp. 423–444. DOI: 10.1016/S0378-4266(01)00228-X.

Levantesi, S. and Menzietti, M. (2012). Managing longevity and disability risks in life annuities with long term care. *Insurance: Mathematics and Economics* 50.3, pp. 391–401. DOI: 10.1016/j.insmatheco.2012.01.004.

Meira-Machado, L. and Roca-Pardinas, J. (2011). p3state.msm: Analyzing Survival Data from an Illness-Death Model. *Journal of Statistical Software* 38.3, pp. 1–18. DOI: 10.18637/jss.v038.i03.

Meira-Machado, L., de Uña-Álvarez, J., and Cadarso-Suárez, C. (2006). Nonparametric estimation of transition probabilities in a non-Markov illness–death model. *Lifetime Data Analysis* 12.3, pp. 325–344. DOI: 10.1007/s10985-006-9009-x.

Meira-Machado, L., de Uña-Álvarez, J., and Datta, S. (2014). Nonparametric estimation of conditional transition probabilities in a non-Markov illness-death model. *Computational Statistics*, pp. 1–21. DOI: 10.1007/s00180-014-0538-6.

Mostajabi, F. and Datta, S. (2013). Nonparametric regression of state occupation, entry, exit, and waiting times with multistate right-censored data. *Statistics in Medicine* 32.17, pp. 3006–3019. DOI: 10.1002/sim.5703.

Peng, L. and Fine, J. P. (2006). Nonparametric estimation with left-truncated semicompeting risks data. *Biometrika* 93.2, pp. 367–383. DOI: 10.1093/biomet/93.2.367.

Pepe, M. S. (1991). Inference for Events With Dependent Risks in Multiple Endpoint Studies. *Journal of the American Statistical Association* 86.415, pp. 770–778. DOI: 10.2307/2290411.

Plisson, M. (2009). Assurabilité et développement de l'assurance dépendance. Ph.D. thesis. Paris: Université Paris Dauphine.

Pritchard, D. J. (2006). Modeling disability in long-term care insurance. *North American Actuarial Journal* 10.4, pp. 48–75. DOI: 10.1080/10920277.2006.10597413.

R Core Team (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Rotolo, F., Legrand, C., and Van Keilegom, I. (2013). A simulation procedure based on copulas to generate clustered multi-state survival data. *Computer Methods and Programs in Biomedicine* 109.3, pp. 305–312. DOI: 10.1016/j.cmpb.2012.09.003.

Salazar, J. C., Schmitt, F. A., Yu, L., Mendiondo, M. M., and Kryscio, R. J. (2007). Shared random effects analysis of multi-state Markov models: application to a longitudinal study of transitions to dementia. *Statistics in medicine* 26.3, pp. 568–580. DOI: 10.1002/sim.2437.

Sánchez-Sellero, C., Manteiga, W. G., and Van Keilegom, I. (2005). Uniform Representation of Product-Limit Integrals with Applications. *Scandinavian Journal of Statistics* 32.4, pp. 563–581. DOI: 10.1111/j.1467-9469.2005.00453.x.

Suzukawa, A. (2002). Asymptotic properties of Aalen-Johansen integrals for competing risks data. *Journal of the Japan Statistical Society* 32, pp. 77–93. DOI: 10.14490/jjss.32.77.

Titman, A. (2015). Transition Probability Estimates for Non-Markov Multi-State Models. *Biometrics*. DOI: 10.1111/biom.12349.

Tomas, J. and Planchet, F. (2013). Multidimensional smoothing by adaptive local kernel-weighted log-likelihood: Application to long-term care insurance. *Insurance: Mathematics and Economics* 52.3, pp. 573–589. DOI: 10.1016/j.insmatheco.2013.03.009.