SORT 48 (2) July-December 2024, 209-234 DOI: 10.57645/20.8080.02.19

Second-order Markov multistate models

Mireia Besalú[1] and Guadalupe Gómez Melis[2]

Abstract

Multistate models are well developed for continuous and discrete times under a first-order Markov assumption. Motivated by a cohort of COVID-19 patients, a multistate model was designed based on 14 transitions among 7 states of a patient. Since a preliminary analysis showed that the first-order Markov condition was not met for some transitions, we have developed a second-order Markov model where the future evolution not only depends on the state at the current time but also on the state at the preceding time. Under a discrete time analysis, assuming homogeneity and that past information is restricted to two consecutive times, we expanded the transition probability matrix and proposed an extension of the Chapman-Kolmogorov equations. We propose two estimators for the second-order transition probabilities and illustrate them within the cohort of COVID-19 patients.

MSC: *62M09, 62N02, 60J10.*

Keywords: *Multistate models, Non-Markov, COVID-19.*

1. Introduction

Multistate models (MSM) provide a very convenient methodology to describe the life history of an individual which at any time occupies one of a few possible states. In particular, they are appropriate to describe the clinical course of a disease and are routinely used in research to model the progression of patients among different states.

MSM theoretical justification is based on the theory of stochastic processes, that is, on sets of random variables representing the evolution of a process over time. The time can be chosen to be discrete or continuous; while discrete times assume a stepwise process where the fixed time between successive steps is not part of the model, continuous

---

[1] Departament Genètica, Microbiologia i Estadística, Universitat de Barcelona, Barcelona, Spain.

[2] Departament d'Estadística i Investigació operativa, Universitat Politècnica de Catalunya, Jordi Girona 1-3, 08034 Barcelona, Spain.

Received: March 2023
Accepted: February 2024

time models allow changes of the states at any time. This class of models allows for an extremely flexible approach that can model almost any kind of longitudinal failure time data. This is particularly relevant for modeling different events, which have an event-related dependence, like the occurrence of a disease changing the risk of death (Hougaard, 1999).

The first-order Markov condition establishes that the future evolution of the stochastic process only depends on the current state and is frequently assumed in multistate models. However, this condition might often be not too realistic to describe clinical situations. To test it, Titman and Putter (2020) develop general log-rank tests that can be applied to general multistate models under right-censoring.

To circumvent the fact that the first-order Markov condition does not hold, the state space could have been extended with new states formed by two adjacent states of the original model. In this case, a first-order Markov condition would probably be satisfied, but the resulting model would be more complex and more difficult to interpret. In addition, the extended model would require more data to be estimated. See, for example, COVID model proposed by Mody et al. (2020), instead of one state of death, they have three depending on the history of the patient.

Another plausible approach to lessen the first-order Markov assumption is to consider a higher-order Markov process. A Markov process of order k is such that the dependence of the process on the whole history is only through the k states previously occupied. Although it is often observed that higher-order Markov processes can model the data better, models for Markov processes of higher order are scarcely used in practice because they depend on a very large number of parameters, leading to computational difficulties (Ching, Fung and Ng, 2003; Logan, 1981). Most instances of higher-order Markov models, which have been used so far, involved discrete time models (known as Markov chains). Tong (1975) defines a k order Markov chain $\{X_1, \cdots, X_n, \cdots\}$ as the one such that the conditional probabilities satisfy

$$P(X_n|X_{n-1}, X_{n-2}, \cdots) = P(X_n|X_{n-1}, X_{n-2}, \cdots, X_{n-k}) \tag{1}$$

for all n, where $k > 0$ is the smallest integer holding the above condition.

In this paper, we propose second-order Markov multistate models as a way of enriching the pathway information and still control the number of parameters while keeping the interpretability of the transition probabilities. Analysis using second-order Markov models are scarce. Among them, Shorrocks (1976) investigated the Markovian assumption in modelling income mobility and concluded that transition rates should depend on both current income and immediate past history, hence, a second-order Markov model was implemented. Shamshad et al. (2005) uses a second-order Markov model for synthetic generation of wind speed time series data.

Second-order Markov models assume that the progression of the individuals not only depends on the current state but also on the state at the preceding time. Second-order Markov multistate models are characterized by means of a $M \times M \times M$ tensor, where M is the number of states. In this work, we define an extended transition probability matrix as M different matrices of order $M \times M$. To be able to compute n-step transition probabilities, we extend the first-order Chapman-Kolmogorov equations. We propose two different estimators for the second-order transition probabilities. We continue the paper with an illustration consisting of a cohort of more than 2000 COVID-19 patients from five hospitals in the Barcelona metropolitan area who were hospitalized during the first wave of the coronavirus pandemic (March-April 2020). For this data we have built a multistate model based on 14 possible transitions among the seven states where a patient can be in after his/her admission. We have adopted a second-order Markov based on plausible medical interpretations and in view of our available data.

We estimate the second-order transition probabilities and based on those we compute, among others, the transition probability from one state to another, after a given number of hospitalized days, and differentiating between patients that arrive to the hospital with severe pneumonia from those who arrive to the hospital with mild pneumonia. The paper ends with a discussion on shortcomings while setting the path for future research.

2. Characterization of first-order Markov multistate processes

A multistate process is a continuous (or discrete)-time stochastic process $X = \{X_t,\ t \geq 0\}$ taking values in a discrete state space $\mathcal{S} = \{1, \cdots, M\}$. We denote by $\mathcal{F}_t := \sigma\{X_s : s \leq t\}$ a σ-algebra consisting on the observation of the process over the interval $[0,t]$ and we refer to it as a filtration. We can think of a filtration as the history of the process up to time t containing the information on the previous occupied states up until time t.

The law of a multistate process is defined by its finite-dimensional distribution and is fully characterized through either one of the following three transition functions: transition probabilities, transition intensities or cumulative transition intensities. The *transition probability* between states h and j for times s and t, $s < t$ is defined by:

$$P_{hj}(s,t;\mathcal{F}_{s-}) = P(X_t = j \mid X_s = h; \mathcal{F}_{s-}) \qquad \text{for} \qquad h,j \in \mathcal{S} = \{1, \cdots, M\}$$

and denotes the probability of the process being at state j at time t knowing that it has been at state h at time s as well as knowing all the previous trajectory before s. The *transition intensity* between states h and j, for time t is defined by:

$$\alpha_{hj}(t;\mathcal{F}_{t-}) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P_{hj}(t, t+\Delta t; \mathcal{F}_{t-}) \qquad \text{for} \qquad h,j \in \mathcal{S} = \{1, \cdots, M\}$$

and denotes the instantaneous probability to change from state h to state j at time t. The *cumulative (integrated) transition intensity* between states h and j at time t is defined by:

$$A_{hj}(t;\mathcal{F}_{t-}) = \int_0^t \alpha_{hj}(u;\mathcal{F}_{u-}) du \qquad \text{for} \qquad h,j \in \mathcal{S} = \{1, \cdots, M\}$$

Transition probabilities, transition intensities and cumulative transition intensities are summarized by means of $M \times M$ matrices. In particular, for every trajectory collected in $\mathcal{F}_{t-}$ and for every s,t such that $s<t$, we denote by $\mathbf{P}$ the transition probability matrix

$$\mathbf{P}(s,t;\mathcal{F}_{s-}) = \{P_{hj}(s,t;\mathcal{F}_{s-}); h,j \in \mathcal{S} = \{1,\cdots,M\}\}.$$

2.1. First-order Markov and homogeneity assumptions

It is clear that some restrictions have to be made in order to estimate the transition probabilities $P_{hj}(s,t;\mathcal{F}_{s-})$ for every pair of states h and j, for every pair of times s and t and for all the possible trajectories before s. The Markov and the homogeneity assumptions are key to make inferences feasible.

Definition 2.1. *A multistate process satisfies the first-order Markov assumption if for all* $h,j \in \mathcal{S} = \{1,\cdots,M\}$ *and* s,t *such that* $s<t$

$$P_{hj}(s,t;\mathcal{F}_{s-}) = P(X_t = j \mid X_s = h;\mathcal{F}_{s-}) = P(X_t = j \mid X_s = h) = P_{hj}(s,t). \qquad (2)$$

That is, under the first-order Markov assumption, different trajectories before s *will not change the transition probabilities. Under the first-order Markov assumption, an* $M\times M$ *matrix,* $P(s,t)$*, is needed for every* (s,t)

$$\mathbf{P}(s,t) = \{P_{hj}(s,t);\ h,j \in \mathcal{S} = \{1,\cdots,M\}\}$$

Definition 2.2. *A first-order Markov multistate process is said to be homogeneous if the transition probability between any states at given times* $t,s \ (s<t)$ *depends only on the difference between these two times* $(t-s)$*, that is,*

$$P_{hj}(s,t) = P_{hj}(0,t-s) = P_{hj}(t-s).$$

In this case only a $M\times M$ *matrix* $\mathbf{P}(t)$ *for every time* t *is needed.*

2.2. Markov Test

We should validate the (first-order) Markov condition if we want to proceed analysing the data under this assumption. One choice would be to include the time of entry into each state as a covariate within a Cox model and test its significance through a likelihood ratio test (Kay, 1986). A second possibility would be to use the stratified version of the Commenges-Andersen's test to detect a shared frailty. Other authors (Rodríguez-Girondo and de Uña Álvarez, 2012) have developed local and global tests for the Markov conditions based upon the observed Kendall's τ for the progressive three-state illness-death model.

In this paper, and in the subsequent COVID-19 analysis, we will validate the Markov assumption for each transition by means of Titman and Putter (2020)'s test that we briefly describe. The main idea of this test is that under the first-order Markov assumption, the rate of transitions at time $t > s$ will not be affected by the state occupied at time s. If we want to check the Markov assumption for the transition between the states l and m ($l, m \in \mathcal{S}$), we divide the subjects into two different groups: the ones that at time s are in a fixed state $j \in \mathcal{S}$ and the ones who are not there. Then, for each transition ($l \to m$) the null hypothesis for a fixed state j and fixed time s ($s \in [t_0, t_{\max}] \subset [0, \tau]$, τ total follow-up) is stated as:

$$\mathrm{H}_{0s}^{(j)}(l,m): \alpha_{lm}(t \mid X(s) = j) = \alpha_{lm}(t \mid X(s) \neq j) \quad \text{for any} \quad t \in [s, \tau]$$

and can be tested with the log-rank statistic

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

where $\delta_i^{(j)}(s) = \mathbb{1}\{X_i(s) = j\}$ denotes whether individual i has been in state j for time s, $Y_i(t)$ is the at risk indicator for the process $X_i(t)$, $Y_{il}(t) = \mathbb{1}\{X_i(t^-) = l\} Y_i(t)$ is the at risk indicator of transition $l \to m$ for subject i and $N_i^{(lm)}(t)$ is the counting process reporting the number of times of the transition $l \to m$ up to time t.

The standardized statistics

$$\overline{U}_s^{(j)}(l,m) = \frac{U_s^{(j)}(l,m)}{\sqrt{\widehat{Var}(U_s^{(j)})(l,m)}}.$$

can be compared to a $N(0,1)$. Moreover, $\left\{\overline{U}_s^{(j)}(l,m), s \in [t_0, t_{\max}]\right\}$ converges to a zero mean Gaussian process with a covariance function that can be consistently estimated.

Given the null hypothesis for a fixed state j

$$\mathrm{H}_0^{(j)}(l,m): \alpha_{lm}(t \mid X(s) = j) = \alpha_{lm}(t \mid X(s) \neq j) \quad \forall s \in [t_0, t_{\max}] \subset [0, \tau] \text{ and } t \in [s, \tau],$$

a global test statistic can be defined based on summary statistics of $\{\overline{U}_s^{(j)}(l,m), s \in [t_0, t_{\max}]\}$ such as $\int_{t_0}^{t_{\max}} \left|\overline{U}_s^{(j)}(l,m)\right| ds$, $\sup_{s \in [t_0, t_{\max}]} |\overline{U}_s^{(j)}(l,m)|$ or $\int_{t_0}^{t_{\max}} w(s) |\overline{U}_s^{(j)}(l,m)| ds$ for some weight function $w(s)$.

Finally, an overall test statistic for the null hypothesis for any possible j and for all $s \in [t_0, t_{\max}] \subset [0, \tau]$ and $t \in [s, \tau]$

$$\mathrm{H}_0(l,m): \alpha_{lm}(t \mid X(s) = j) = \alpha_{lm}(t \mid X(s) \neq j) \quad \forall j, \forall s \in [t_0, t_{\max}] \subset [0, \tau] \text{ and } \forall t \in [s, \tau]$$

can be defined from the global test statistics, for instance as the mean, the maximum or weighted mean of them. These tests are implemented in `R` with the function `MarkovTest` of the package `mstate` of de Wreede, Fiocco and Putter (2011). Details of how are implemented are postponed to the illustration in Section 6.3.

3. Characterization of second-order Markov transition probabilities

To address the limitations of the first-order Markov assumption, we introduce a second-order Markov assumption, which acknowledges that the future evolution of the stochastic process depends not only on its current state but also on the state it occupied in the preceding time. We begin by defining the second-order Markov transition probabilities and describing how they can be summarized into a set of as many matrices as states.

Definition 3.1. *For times* (s,t,u), $s<t<u$ *and states* h,j,k, *the probability* $P_{hjk}(s,t,u;\mathcal{F}_{s-}) = P(X_u = k \mid X_s = h, X_t = j; \mathcal{F}_{s-})$ *satisfies a second-order Markov assumption if and only if*

$$\begin{aligned} P_{hjk}(s,t,u;\mathcal{F}_{s-}) &= P(X_u = k \mid X_s = h, X_t = j | \mathcal{F}_{s-}) \\ &= P(X_u = k \mid X_s = h, X_t = j) = P_{hjk}(s,t,u). \end{aligned}$$

Under the second-order Markov assumption the transition probabilities are summarized, for every three times (s,t,u), $s<t<u$, by an $M \times M \times M$ tensor $\mathbf{P}(s,t,u)$

$$\mathbf{P}(s,t,u) = \{P_{hjk}(s,t,u); h,j,k \in \mathcal{S} = \{1,\cdots,M\}\}.$$

In order to have a more manageable mathematical object we denote, for each state $h \in \mathcal{S}$, a matrix of dimension M, $\mathbf{P}_{(\mathbf{h})}(s,t,u)$ as follows:

$$\mathbf{P}_{(\mathbf{h})}(s,t,u) = \left(P_{hjk}(s,t,u)\right)_{j,k\in\mathcal{S}},$$

hence, the tensor $\mathbf{P}(s,t,u)$ of transition probabilities can be equivalently represented as M matrices of order M for each $s<t<u$.

Remark 3.2. *The matrices* $\mathbf{P}_{(\mathbf{h})}(s,t,u)$ *are not always stochastic matrices because*

$$\sum_{k\in\mathcal{S}} P_{hjk}(s,t,u) = \begin{cases} 0 & \textit{if } \forall j,k, \textit{ the transitions } h \to j \textit{ or } j \to k \textit{ are not possible} \\ 1 & \textit{otherwise.} \end{cases}$$

For example, if h is an absorbent state all the matrices will be 0 except for the element $P_{hhh} = 1$.

Definition 3.3. *A second-order Markov multistate process is said to be homogeneous if the transition probability between any three states at given times* (s,t,u), $s<t<u$, *depends only on the differences* $t-s$ *and* $u-t$ *between the two consecutive times that is,*

$$P_{hjk}(s,t,u) = P_{hjk}(t-s,u-t)$$

In this case only a $M \times M \times M$ *tensor* $\mathbf{P}(s,t)$ *for every pair of times* (s,t) $(s < t)$ *is needed*

$$\mathbf{P}(s,t) = \{P_{hjk}(s,t);\, h,j,k \in \mathcal{S} = \{1,\cdots,M\}\}.$$

Using the previous notation, denote as $\mathbf{P}_{(\mathbf{h})}(\mathbf{s},\mathbf{t})$ *the matrix of dimension M for each state* $h \in \mathcal{S}$ *and for every pair of times* (s,t) $(s < t)$*, that is,*

$$\mathbf{P}_{(\mathbf{h})}(s,t) = \left(P_{hjk}(s,t)\right)_{j,k\in\mathcal{S}}.$$

Note that the tensor $\mathbf{P}(s,t)$ of transition probabilities under homogeneity can be equivalently represented as M matrices of dimension M for each two times $(s,t); s < t$ where s stands for the time from h to j and t stands for the time from j to k.

4. Computation of transition probabilities

Given that clinical outcomes are often collected in days and aiming to compute the probability of being in a given state after a certain number of days, we consider in this section a discrete-time multistate process instead of a continuous-time stochastic process defined for $t \in [0,T]$. Other instances of discrete-time multistate process have been used to model COVID-19 disease progression and clinical outcomes (Chakladar et al., 2022).

In this section we provide the expressions to compute general probabilities, such as

$$P_{hjl}(s,s+n,s+n+m) = P(X_{s+n+m} = l \mid X_{s+n} = h, X_s = j). \tag{3}$$

for any three states j,h,l at any three times $s, s+n, s+n+m$. In order to get there we start extending the Chapman-Kolmogorov equations from first to second-order Markov chains based on 2-step second-order transition probabilities, that is, on $P(X_{s+2} = l \mid X_{s-1} = h, X_s = j)$. In Subsection 4.3, probabilities (3) are written as a function of the 1-step first-order transition probabilities, $P(X_{s+1} = h|X_s = j)$, and 1-step second-order transition probabilities for consecutive times, $P(X_{s+n+1} = l \mid X_{s+n} = h, X_{s+n-1} = j)$.

Remark 4.1. *We use the term n-step* ***second-order*** *transition probabilities to refer to n-step transition probabilities conditioned to 2 consecutive times, that is,* $P(X_{s+n} = l \mid X_{s-1} = h, X_s = j)$*, for* $n \geq 1$*. We also define the n-step* ***first-order*** *transition probabilities as follows* $P(X_{s+n} = l \mid X_s = h)$ *for* $n \geq 1$*.*

4.1. Chapman-Kolmogorov equations for first-order Markov chains

A first-order discrete-time multistate models, known as Markov chain, taking values in a discrete state space $\mathcal{S} = \{1,\cdots,M\}$ is the discrete version of a first-order continuous-time Markov process. Hence, a Markov chain is a stochastic model describing a sequence of possible events happening on discrete times in which the probability of each event depends only on the state attained in the previous event. The Chapman-Kolmogorov relation is an important result in the theory of (discrete) Markov chains as it provides a method for calculating the n-step transition probabilities.

The Chapman-Kolmogorov equations, for any $s, t, u \in \mathbb{N}$ $(s < u < t)$ and any two states h, j are given by

$$P_{hj}(s,t) = \sum_{l=1}^{m} P_{hl}(s,u) P_{lj}(u,t), \tag{4}$$

where $P_{hj}(s,t) = P(X_t = j \mid X_s = h)$ is the transition probability defined in (2). Chapman-Kolmogorov equations follow as a consequence of the Markov condition. Chapman-Kolmogorov equations allow to reduce the general computation of $P_{hj}(s,t)$, for any $s < t$, $(s,t \in \mathbb{N})$ to the computation of 1-step first-order transition probabilities, $P_{hj}(s,s+1)$, that is,

$$P_{hj}(s,s+n) = \sum_{l=1}^{M} P_{hl}(s,s+1) P_{lj}(s+1,s+n), \tag{5}$$

Denote by $\mathbf{P}(s)$ the one-time step transition probability matrix under the Markov assumption, that is,

$$\mathbf{P}(s) = \{P_{hj}(s);\, h,j \in \mathcal{S} = \{1,\cdots,M\}\}$$

where $P_{hj}(s)$ stands for $P_{hj}(s,s+1)$. The collection of matrices $\mathbf{P}(s)$ is reduced to the transition probability matrix $\mathbf{P}$ given by

$$\mathbf{P} = \{P_{hj} = P_{hj}(1);\, h,j \in \mathcal{S} = \{1,\cdots,M\}\}$$

under the homogeneity assumption (see Definition 2.2). Hence, to study the evolution of the process for more than one time step, and thanks to the Chapman-Kolmogorov equations (4), it is only necessary to calculate the one-time initial transition probabilities. In the next section we develop an extension of this result for second-order Markov chains.

4.2. Chapman-Kolmogorov equations for second-order Markov chains

Under second-order Markov and homogeneity assumptions the transition probability matrices defined in Section 3 satisfy, for times $s < t < u$ and states h, j, k:

$$P_{hjk}(s,t,u;\mathcal{F}_{s-}) = P(X_u = k \mid X_s = h, X_t = j) = P_{hjk}(s,t,u) = P_{hjk}(t-s,u-t)$$

In particular, for any $s \in \mathbb{N}$, $s > 1$ and consecutive times $s, s+1, s+2$, computation of the probabilities $P_{hjk}(s,s+1,s+2;\mathcal{F}_{s-})$ is reduced to the computation of 1-step second-order transition probabilities, that is,

$$\begin{aligned} P_{hjk}(s,s+1,s+2;\mathcal{F}_{s-}) &= P(X_{s+2} = k \mid X_s = h, X_{s+1} = j) \\ &= P_{hjk}(s,s+1,s+2) = P_{hjk}(1,1). \end{aligned}$$

Next theorem presents the equations to compute n-step second-order transition probabilities such as

$$P(X_{s+n+1} = l | X_{s+1} = j, X_s = h) = P_{hjl}(1,n), \tag{6}$$

for $s, n \in \mathbb{N}, n > 1$ and $l, j, k \in \mathcal{S}$ using only the initial transition probabilities $P(X_3 = l|X_2 = j, X_1 = h) = P_{hjl}(1,1) = P_{hjl}$.

Since our desired probabilities only depend on n, without lost of generality we can assume $s = 1$ and the probabilities at (6) can be equivalently written as

$$P(X_{n+2} = l|X_2 = j, X_1 = h), \quad \text{for} \quad n \in \mathbb{N},\ l, j, k \in \mathcal{S}.$$

Recall that these transition probabilities can be summarized into M matrices of dimension M for each $n \in \mathbb{N}$. So for each state $h \in \mathcal{S}$ and $n > 1$

$$\mathbf{P}_{(\mathbf{h})}(1,2,n+2) = \left(P_{hjl}(1,n)\right)_{j,k\in\mathcal{S}} = \left(P_{hjl}(n)\right)_{j,l\in\mathcal{S}},$$

we will omit n of the previous notation when $n = 1$.

Notation 4.2. *Previous to the main result, we will present the matrix notation used in order to simplify the reading.*

- *Row j of matrix h:* $P_{j\cdot(h)} = (P_{hjk})_{k\in\mathcal{S}}$.
- *Column k of matrix h:* $P_{\cdot k(h)} = (P_{hjk})_{j\in\mathcal{S}}$.
- *We will use the* $*$ *symbol to denote that the elements of the row multiply each row of the matrix. For example* $P_{hj\cdot} * P_{(h)}$ *means that element* P_{hjk} *multiplies all the elements of the row k of* $P_{(h)}$.
- $\mathbf{P}^{(l)}$ *is the matrix composed with the l column of each of the* P_h *matrices* $h \in \mathcal{S}$.
- Tr() *will denote the trace of a matrix.*

Theorem 4.3. *Assume* $(X_n)_{n\in\mathbb{N}}$ *is an homogeneous second-order Markov chain. For any states* $h, j, l \in \mathcal{S}$ *and the notation defined in Notation 4.2, we have (where* Tr *denotes trace)*

$$P(X_4 = l \mid X_2 = j, X_1 = h) = \mathbf{P}_{j\cdot(h)} \cdot \mathbf{P}_{\cdot l(j)} \tag{7}$$

$$P(X_5 = l \mid X_2 = j, X_1 = h) = \mathrm{Tr}\left(P_{hj\cdot} * \mathbf{P}_{(j)} \cdot \mathbf{P}^{(l)}\right) \tag{8}$$

$$P(X_6 = l \mid X_2 = j, X_1 = h) = \sum_{k_1=1}^{M} P_{hjk_1} \cdot \mathrm{Tr}\left(P_{jk_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)}\right) \tag{9}$$

$$P(X_7 = l \mid X_2 = j, X_1 = h) = \sum_{k_2=1}^{M} \sum_{k_1=1}^{M} P_{hjk_2} P_{jk_2k_1} \cdot \mathrm{Tr}\left(P_{k_2k_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)}\right) \tag{10}$$

General case $n \geq 7$

$$P(X_{n+1} = l \mid X_2 = j, X_1 = h) = \sum_{k_{n-4}=1}^{M} \cdots \sum_{k_2=1}^{M} \sum_{k_1=1}^{M} P_{hjk_{n-4}} P_{jk_{n-4}k_{n-3}} \cdots P_{k_3k_2k_1} \times \mathrm{Tr}\left(P_{k_2k_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)}\right)$$

Proof. The proof will be divided into three steps.

Step 1. We are proving the first case (7).

Using the total probabilities Theorem and the second-order Markov property,

$$\begin{aligned}
P(X_4 = l \mid X_2 = j, X_1 = h) &= \sum_{k=1}^{M} P(X_4 = l, X_3 = k \mid X_2 = j, X_1 = h) \\
&= \sum_{k=1}^{M} P(X_3 = k \mid X_2 = j, X_1 = h) P(X_4 = l \mid X_3 = k, X_2 = j, X_1 = h) \\
&= \sum_{k=1}^{M} P(X_3 = k \mid X_2 = j, X_1 = h) P(X_4 = l \mid X_3 = k, X_2 = j) \\
&= \sum_{k=1}^{M} P_{hjk} \cdot P_{jkl} \qquad (11)
\end{aligned}$$

We recall that $P(X_4 = l \mid X_3 = k, X_2 = j) = P_{jkl}(1,1) = P_{jkl}$ since for all the assumptions only the difference between times determine the transition probabilities. From here we can write it in matricial form as in (7).

Step 2. Now we focus in the second case (8).

Using the same argument of the previous case and also the result obtained (11)

$$\begin{aligned}
P(X_5 = l \mid X_2 = j, X_1 = h) &= \sum_{k_1=1}^{M} P(X_5 = l, X_3 = k_1 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_1=1}^{M} P(X_5 = l \mid X_3 = k_1, X_2 = j, X_1 = i) P(X_3 = k_1 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_1=1}^{M} P(X_5 = l \mid X_3 = k_1, X_2 = j) P(X_3 = k_1 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_1=1}^{M} \left(\sum_{k_2=1}^{M} P_{k_1 k_2 l} \cdot P_{j k_1 k_2} \right) \cdot P_{hjk_1}
\end{aligned}$$

Now, if we want to write it in a matricial way we can observe that $\sum_{k_2=1}^{M} P_{k_1 k_2 l} \cdot P_{j k_1 k_2}$ corresponds to the product of the l column of each matrix by the matrix $\mathbf{P}_{(j)}$ and then each row of this matrix product is multiplied by the probabilities P_{hjk_1} which are the elements of row j of matrix $\mathbf{P}_{(h)}$. From here we obtain the formula in (8).

Step 3. We follow by proving (9).

We repeat here the arguments in the previous steps and also we apply the previous results.

$$\begin{aligned}
P(X_6 = l \mid X_2 = j, X_1 = h) &= \sum_{k_3=1}^{M} P(X_6 = l, X_3 = k_3 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_3=1}^{M} P(X_6 = l \mid X_3 = k_3, X_2 = j, X_1 = h) P(X_3 = k_3 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_3=1}^{M} P(X_6 = l \mid X_3 = k_3, X_2 = j) P(X_3 = k_3 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_3=1}^{M} P_{hjk_3} \left(\sum_{k_1=1}^{M} P_{jk_3k_1} \sum_{k_2=1}^{M} P_{k_1k_2l} \cdot P_{k_3k_1k_2} \right) \\
&= P_{hj1} \left(\sum_{k_1=1}^{M} P_{jk_3k_1} \sum_{k_2=1}^{M} P_{k_1k_2l} \cdot P_{k_3k_1k_2} \right) + \ldots + P_{hjM} \left(\sum_{k_1=1}^{M} P_{jk_3k_1} \sum_{k_2=1}^{M} P_{k_1k_2l} \cdot P_{k_3k_1k_2} \right)
\end{aligned}$$

We observe that we obtain a similar expression as the previous step but multiplied by the probabilities $P_{hj\cdot}$ that correspond to the row j of matrix $\mathbf{P}_{(h)}$.Thus, the matricial expression for this case is immediate.

From here, in order to prove Equation (10) and the general case we can just repeat the same arguments as in this last case to easily obtain the general formula by induction. In these two cases the principal modifications of the matricial form will focus in adding one sum for each step. ■

Corollary 4.3.1. *Assume $(X_n)_{n\in\mathbb{N}}$ is an homogeneous second-order Markov chain. For time $s > 0$, any states $h, j, l \in \mathcal{S}$ and the notation defined in Notation 4.2, we have*

$$\begin{aligned}
P(X_{s+3} = l \mid X_{s+1} = j, X_s = h) &= \mathbf{P}_{j\cdot(h)} \cdot \mathbf{P}_{\cdot l(j)} \\
P(X_{s+4} = l \mid X_{s+1} = j, X_s = h) &= \mathrm{Tr}\left(P_{hj\cdot} * \mathbf{P}_{(j)} \cdot \mathbf{P}^{(l)} \right) \\
P(X_{s+5} = l \mid X_{s+1} = j, X_s = h) &= \sum_{k_1=1}^{M} P_{hjk_1} \cdot \mathrm{Tr}\left(P_{jk_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)} \right) \\
P(X_{s+6} = l \mid X_{s+1} = j, X_s = h) &= \sum_{k_2=1}^{M} \sum_{k_1=1}^{M} P_{hjk_2} P_{jk_2k_1} \cdot \mathrm{Tr}\left(P_{k_2k_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)} \right)
\end{aligned}$$

General case $n \geq 7$

$$\begin{aligned}
P(X_{s+n} = l \mid X_{s+1} = j, X_s = h) = \sum_{k_{n-4}=1}^{M} \ldots \sum_{k_2=1}^{M} \sum_{k_1=1}^{M} & P_{hjk_{n-4}} P_{jk_{n-4}k_{n-3}} \ldots P_{k_3k_2k_1} \\
& \times \mathrm{Tr}\left(P_{k_2k_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)} \right) \quad (12)
\end{aligned}$$

Remark 4.4. *We observe that the extended Chapman-Kolmogorov equations only consider the case where the two past times are consecutive.*

4.3. Computation of arbitrary transition probabilities

So far, the extended Chapman-Kolmogorov equations have only considered those cases where the two past times were consecutive. For some models and specific cases with non return states it is possible to compute these probabilities.

In this section we prove that for any 3 times $s, s+n, s+n+m$ ($s > 0$, $n,m > 1$) transition probabilities defined as follows:

$$P_{hjl}(s,s+n,s+n+m) = P(X_{s+n+m} = l \mid X_{s+n} = h, X_s = j)$$

when the two past times are not consecutive, can be written as a function of the 1-step first-order transition probabilities and the 1-step second-order transition probabilities for consecutive times. The tools presented in the previous subsections will be now crucial to obtain the desired probability.

Theorem 4.5. *Assume* $(X_n)_{n\in\mathbb{N}}$ *is an homogeneous second-order Markov chain. For any states* $h,j,l \in \mathcal{S}$ *and time* $s > 0$ *and* $n,m > 1$ *we have*

$$P(X_{s+n+m} = l \mid X_{s+n} = h, X_s = j) = \frac{\sum_{e_1=1}^{M} \cdots \sum_{e_{n-1}=1}^{M} P(X_{s+n+m} = l \mid X_{s+n} = h, X_{s+n-1} = e_{n-1}) P_{e_{n-1}e_{n-2}h}(s+n) \ldots P_{je_1e_2}(s+2) P(X_{s+1} = e_1 \mid X_s = j)}{P(X_{s+n} = h \mid X_s = j)}$$

where $P(X_{s+n+m} = l \mid X_{s+n} = h, X_{s+n-1} = e_{n-1})$ *can be written in terms of* 1*-step second-order transition probabilities for consecutive times as it is shown in* (12) *and* $P(X_{s+n} = h \mid X_s = j)$ *can be written in terms of* 1*-step first-order transition probabilities as it is shown in* (5).

Proof. Indeed,

$$\begin{aligned} P_{hjl}(s,s+n,s+n+m) &= P(X_{s+n+m} = l \mid X_{s+n} = h, X_s = j) \\ &= \frac{P(X_{s+n+m} = l, X_{s+n} = h, X_s = j)}{P(X_{s+n} = h, X_s = j)} \end{aligned}$$

We can write the numerator in the following way

$$\begin{aligned} &P(X_{s+n+m} = l, X_{s+n} = h, X_s = j) \\ &\quad = \sum_{e_1=1}^{M} \cdots \sum_{e_{n-1}=1}^{M} P(X_{s+n+m} = l, X_{s+n} = h, X_{s+n-1} = e_{n-1}, \ldots, X_{s+1} = e_1, X_s = j) \\ &\quad = \sum_{e_1=1}^{M} \cdots \sum_{e_{n-1}=1}^{M} P(X_{s+n+m} = l \mid X_{s+n} = h, X_{s+n-1} = e_{n-1}, \ldots, X_{s+1} = e_1, X_s = j) \\ &\qquad\qquad \times P(X_{s+n} = h, X_{s+n-1} = e_{n-1}, \ldots, X_{s+1} = e_1, X_s = j) \end{aligned}$$

We can iterate the process and apply the second-order Markov hypothesis and finally obtain

$$
\begin{aligned}
&P(X_{s+n+m}=l,X_{s+n}=h,X_s=j)\\
&\quad=\sum_{e_1=1}^{M}\cdots\sum_{e_{n-1}=1}^{M}P(X_{s+n+m}=l\mid X_{s+n}=h,X_{s+n-1}=e_{n-1})\\
&\qquad\quad\times P(X_{s+n}=h\mid X_{s+n-1}=e_{n-1},X_{s+n-2}=e_{n-2})\times\dots\\
&\qquad\quad\times P(X_{s+2}=e_2\mid X_{s+1}=e_1,X_s=j)\times P(X_{s+1}=e_1|X_s=j)P(X_s=j)
\end{aligned}
$$

For the denominator we just recall the conditional probability definition

$$P(X_{s+n}=h,X_s=j)=P(X_{s+n}=h|X_s=j)P(X_s=j).$$

Now we have already prove the result since $P(X_s=j)$ is in both numerator and denominator and we can simplify it. ■

Remark 4.6. *With the result of this Theorem, the Chapman-Kolmogorov equations for first and second-order and the one-step transition probabilities also for first and second-order we can now compute any transition probability for any triplet of times.*

5. Estimation and inference under second-order Markov assumption

Given three different states $h,\,j,\,l\in\mathcal{S}$ such that $h,\,j$ are not absorbent, the purpose of this Section is to estimate the r-step transition probabilities $P(X_{s+r}=l\mid X_{s-1}=j,X_{s-2}=h)$ for any $s,r\in\mathbb{N}$, $s,r>1$. Under the homogeneity assumption we have that $P(X_s=l\mid X_{s-1}=j,X_{s-2}=h)=P_{hjl}(1,1)=P_{hjl}$ and, following Corollary 4.3.1, in order to estimate $P(X_{s+r}=l\mid X_{s-1}=j,X_{s-2}=h)$ is enough to estimate the initial transition probabilities $P_{hjl}(1,1)=P_{hjl}$.

We assume that individuals are followed until a maximum of T units of time (days as in the illustration). Let $\{X_s^i,\ s=0,1,\cdots,T\}$ denote the non-reversible multistate process for subject $i=1,\dots,n$, where $X_s^i\in\mathcal{S}$. For $i=1,\dots,n$, $s=2,\cdots,T$ and $h,\,j,\,l\in\mathcal{S}$ we define the counting processes

$$N_{hjl}^i(s)=\mathbb{1}\{X_{s-2}^i=h,\,X_{s-1}^i=j,\,X_s^i=l\}$$

counting 1 if subject i has transit from state h to state j and to state l at times $s-2$, $s-1$, s, respectively; and 0 otherwise. The total number of individuals who have followed the path $h\to j\to l$ at times $s-2,s-1,s$ is given by the sum $\widetilde{N}_{hjl}(s)=\sum_{i=1}^{n}N_{hjl}^i(s)$. $\widetilde{N}_{hjl}(s)$ is a binomial random variable with parameters $(n,\pi_{hjl}(s))$ where the probability

$\pi_{hjl}(s)$ corresponds to $\pi_{hjl}(s) = P(X_{s-2} = h, X_{s-1} = j, X_s = l)$. We also define the at-risk process of subject i corresponding to states h and j at times $s = 2, \cdots, T$,

$$Y^i_{hj}(s-1) = \mathbb{1}\{X^i_{s-2} = h, X^i_{s-1} = j\}$$

counting 1 if subject i was at risk of moving to adjacent states to j or stay at j given that he/she was in states h and j at times $s-2$ and $s-1$, respectively. The total number of individuals at risk at time s is given by $\widetilde{Y}_{hj}(s-1) = \sum_{i=1}^n Y^i_{hj}(s-1)$ and corresponds to a binomial random variable with parameters $(n, \pi_{hj}(s-1))$ where $\pi_{hj}(s-1) = P(X_{s-2} = h, X_{s-1} = j)$.

Regarding the estimation of the transition probability $P_{hjl}(s) = P(X_s = l \mid X_{s-1} = j, X_{s-2} = h)$ for a given $s \geq 2$, we will proceed in two different ways. The first one takes advantage of the ratio of the two probabilities $P_{hjl}(s) = \pi_{hjl}(s)/\pi_{hj}(s-1)$ while the second one exploits directly the estimation of the conditional probability $P_{hjl}(s)$.

From a practical point of view we will have to guarantee that the number of individuals at risk $\widetilde{Y}_{hj}(s-1)$ is large enough for a meaningful estimation of P_{hjl}.

5.1. Estimation of P_{hjl} via the Bernoulli probabilities π_{hjl} and π_{hj}

Given that for all $s = 2, \cdots, T$

$$\begin{aligned} P_{hjl} = P_{hjl}(s) = P(X_s = l \mid X_{s-1} = j, X_{s-2} = h) &= \frac{P(X_{s-2} = h, X_{s-1} = j, X_s = l)}{P(X_{s-2} = h, X_{s-1} = j)} \\ &= \frac{\pi_{hjl}(s)}{\pi_{hj}(s-1)} \end{aligned}$$

a natural estimator for P_{hjl} can be built estimating separately both numerator and denominator by $\sum_{s=2}^T \widetilde{N}_{hjl}(s)/n$ and $\sum_{s=2}^T \widetilde{Y}_{hj}(s-1)/n$, respectively. Observe that $\sum_{s=2}^T \widetilde{N}_{hjl}(s)$ corresponds to the total number of individuals that have followed the path $h \to j \to l$ at any three times $(s-2, s-1, s)$ and $\sum_{s=2}^T \widetilde{Y}_{hj}(s-1)$ is the total number of individuals that have followed the path $h \to j$ consecutively at any two times $(s-2, s-1)$.

Definition 5.1. *For given states (h, j, l), the statistic*

$$\widetilde{P}_{hjl} = \frac{\sum_{s=2}^T \widetilde{N}_{hjl}(s)}{\sum_{s=2}^T \widetilde{Y}_{hj}(s-1)} \tag{13}$$

estimates P_{hjl}. Whenever the denominator $\sum_{s=2}^T \widetilde{Y}_{hj}(s-1)$ is equal to 0, meaning that no individuals have contributed to the path $h \to j$, we will take $\widetilde{P}_{hjl} = 0$.

Theorem 5.2. *For given states (h, j, l), the statistic $\widetilde{P}_{hjl}$ defined in (13) is a consistent estimator of P_{hjl}.*

Proof. Note that by the Law of Large Numbers we have convergence in probability of the following two estimators:

$$\frac{1}{n}\sum_{s=2}^{T}\widetilde{N}_{hjl}(s)=\frac{1}{n}\sum_{i=1}^{n}\sum_{s=2}^{T}N^{i}_{hjl}(s)\xrightarrow[n\to\infty]{\mathrm{P}}E\left(\sum_{s=2}^{T}N^{i}_{hjl}(s)\right)=\sum_{s=2}^{T}E\left(N^{i}_{jhl}(s)\right)$$
$$=\sum_{s=2}^{T}\pi_{hjl}(s)$$

$$\frac{1}{n}\sum_{s=2}^{T}\widetilde{Y}_{hj}(s-1)=\frac{1}{n}\sum_{i=1}^{n}\sum_{s=2}^{T}Y^{i}_{hj}(s-1)\xrightarrow[n\to\infty]{\mathrm{P}}E\left(\sum_{s=2}^{T}Y^{i}_{hj}(s-1)\right)$$
$$=\sum_{s=2}^{T}E\left(Y^{i}_{jh}(s-1)\right)=\sum_{s=2}^{T}\pi_{hj}(s-1)$$

Second-order homogeneity implies that

$$\begin{aligned}\pi_{hjl}(s)&=P\{X_{s-2}=h,X_{s-1}=j,X_s=l\}\\&=P\{X_s=l\mid X_{s-2}=h,X_{s-1}=j\}P\{X_{s-1}=j,X_{s-2}=h\}\\&=P_{hjl}(1,1)P\{X_{s-1}=j,X_{s-2}=h\}=P_{hjl}(1,1)\pi_{hj}(s-1),\end{aligned}$$

and we conclude that $\widetilde{P_{jhl}}$ converges in probability to P_{hjl}:

$$\widetilde{P}_{jhl}=\frac{\sum_{s=2}^{T}\tilde{N}_{hjl}(s)}{\sum_{s=2}^{T}\tilde{Y}_{hj}(s-1)}=\frac{\sum_{s=2}^{T}\tilde{N}_{hjl}(s)/n}{\sum_{s=2}^{T}\tilde{Y}_{hj}(s-1)/n}\xrightarrow[n\to\infty]{\mathrm{P}}\frac{\sum_{s=2}^{T}\pi_{hjl}(s)}{\sum_{s=2}^{T}\pi_{hj}(s-1)}$$

$$=\frac{\sum_{s=2}^{T}P_{hjl}(1,1)\pi_{hj}(s-1)}{\sum_{s=2}^{T}\pi_{hj}(s-1)}=P_{hjl}$$

■

5.2. Estimation of P_{hjl} via the conditional probability

For every $s\geq 2$, the relative frequency given by the ratio $\widetilde{N}_{hjl}(s)/\widetilde{Y}_{hj}(s-1)$ is, whenever $\widetilde{Y}_{hj}(s-1)>0$, an straightforward estimator of $P_{hjl}(s)$.

Because the homogeneity assumption, we have for all $s=2,\cdots,T$, $P_{hjl}(s)=P_{hjl}$ and hence, an estimator for P_{hjl} can be obtained as the average of $\widehat{P}_{hjl}(s)$.

Definition 5.3. *For given states* (h,j,l) *and for every* $s\geq 2$, *define the statistic*

$$\widehat{P}_{hjl}(s)=J_{hj}(s-1)\frac{\widetilde{N}_{hjl}(s)}{\widetilde{Y}_{hj}(s-1)}$$

where $J_{hj}(s-1) = \mathbb{1}\{\widetilde{Y}_{hj}(s-1) > 0\}$. *To estimate* P_{hjl} *we define an alternative estimator as follows:*

$$\widehat{P}_{hjl} = \frac{1}{t_{hj}} \sum_{s=2}^{T} \widehat{P}_{hjl}(s). \tag{14}$$

where $t_{hj} = \sum_{s=2}^{T} J_{hj}(s-1)$ *counts the number of times where at least there is an individual experiencing the path* $j \to h$.

Theorem 5.4. *For given states* (h, j, l) *the statistic* $\widehat{P}_{hjl}$ *given in* (14) *is an unbiased estimator of* P_{hjl}.

Proof. We assume that t_{hj}, the number of times where at least there is an individual at risk, is fixed. Then,

$$\begin{aligned}
\mathrm{E}[\widehat{P}_{hjl}] &= \mathrm{E}\left[\frac{1}{t_{hj}} \sum_{s=2}^{T} J_{hj}(s-1) \frac{\widetilde{N}_{hjl}(s)}{\widetilde{Y}_{hj}(s-1)}\right] = \frac{1}{t_{hj}} \sum_{s=2}^{T} \mathrm{E}\left[J_{hj}(s-1) \frac{\widetilde{N}_{hjl}(s)}{\widetilde{Y}_{hj}(s-1)}\right] \\
&= \frac{1}{t_{hj}} \sum_{s=2}^{T} \mathrm{E}\left[\mathrm{E}\left[J_{hj}(s-1) \frac{\widetilde{N}_{hjl}(s)}{\widetilde{Y}_{hj}(s-1)} \,\middle|\, \tilde{Y}_{hj}(s-1)\right]\right] \\
&= \frac{1}{t_{hj}} \sum_{s=2}^{T} \mathrm{E}\left[J_{hj}(s-1) \frac{\mathrm{E}\left[\widetilde{N}_{hjl}(s) \,\middle|\, \tilde{Y}_{hj}(s-1)\right]}{\widetilde{Y}_{hj}(s-1)}\right] \\
&= \frac{1}{t_{hj}} \sum_{s=2}^{T} \mathrm{E}\left[J_{hj}(s-1) \frac{\widetilde{Y}_{hj}(s-1) p_{hjl}}{\widetilde{Y}_{hj}(s-1)}\right] = \frac{1}{t_{hj}} \sum_{s=2}^{T} \mathrm{E}\left[J_{hj}(s-1) p_{hjl}\right] \\
&= p_{hjl} \mathrm{E}\left[\frac{1}{t_{hj}} \sum_{s=2}^{T} J_{hj}(s-1)\right] = p_{hjl}
\end{aligned}$$

where we have used that $\tilde{N}_{hjl}(s) | \tilde{Y}_{hj}(s-1) \sim \mathrm{Bin}(\tilde{Y}_{hj}(s-1), p_{hjl})$ ■

6. DIVINE model

6.1. Description

The dataset we use as illustration corresponds to a cohort of 2076 COVID-19 hospitalised patients (during the first wave of the pandemic, March-April 2020) in five hospitals located in the southern Barcelona metropolitan area (Spain). Since all the patients were monitored until discharge from hospital or death, the transition times (in days) are known exactly for all subjects and there are not incomplete data due to lost to follow-up.

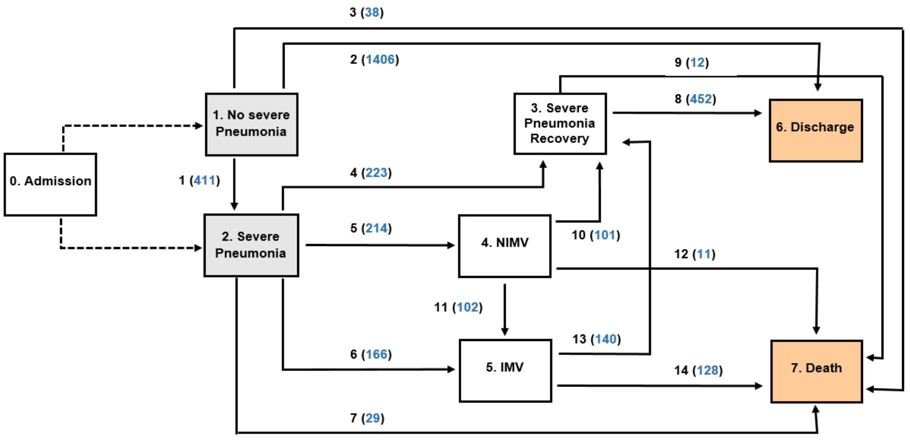

Figure 1. *Graphical representation of the multistate model for modelling the trajectory of hospitalized COVID-19 patients. Seven states are considered and 14 possible transitions (in parentheses the sample size of each transition). NIMV: Non-Invasive mechanical ventilation, IMV: Invasive mechanical ventilation.*

This data is part of the DIVINE project (https://grbio.upc.edu/en/research/highlighted-projects) for which a multidisciplinary research team integrated by researchers from the GRBIO (UPC-UB), Bellvitge University Hospital, and Bellvitge Biomedical Research Institute has collaborated to define a statistical framework with a clear clinician focus on achieving deeper understanding of the severe form of the disease caused by the SARS-CoV-2 virus. Based on the team cooperative knowledge a multistate model with seven states and 14 transitions has been built (see Figure 1 where the numbers in parentheses denote the patients doing that transition). As seen in Figure 1, 7 states are considered: (1) No Severe Pneumonia (NSP), (2) Severe Pneumonia (SP), (3) Severe Pneumonia Recovery (Recov), (4) Non invasive mechanical ventilation (NIMV), (5) Invasive mechanical ventilation (IMV), (6) Discharge (Disch) and (7) Death (Death).

The following considerations are in place:

1. Once a patient has been admitted (state 0), he/she is immediately assigned to one of the two initial states: No Severe Pneumonia and Severe Pneumonia. It is assumed that the process starts at time $t = 0$ in one of these two states.
2. Discharge and Death are absorbing states implying that once a patient has been discharged or has died he/she cannot re-enter to be hospitalised again.
3. The time scale used in this model is days since the hospital admission. For all the transitions, the transition times (in days) are exactly known.
4. Patients can only jump to a neighboring state in a single day.

For more details on the data and the clinical patient characteristics see Pallarès et al. (2023); Garmendia, Cortés and Gómez Melis (2023); Piulachs et al. (2023).

The main goal with this illustration is to study the evolution of the patients without the restriction of a first-order Markov assumption. To do so we start validating for which transitions of the previous multistate model the Markov assumption holds. Next, we will estimate the transition probabilities between two states taking into account that they might depend as well on the immediate previous state. Finally, we will compare the evolution of those patients admitted with No Severe Pneumonia versus those admitted with Severe Pneumonia.

6.2. Description of direct and two-step transitions

Table 1 summarises the number of patients for each direct transition and the number of patients for the corresponding related 2-step transitions (consecutive states but not necessarily consecutive times). For instance, individuals doing the direct transition Recov → Disch might arrive from SP : SP → Recov → Disch from NIMV : NIMV → Recov → Disch or from IMV : IMV → Recov → Disch. Note that we are only considering those direct transitions $j \rightarrow l$ for which there exists, at least, a state k adjacent to j ($k \rightarrow j \rightarrow l$).

Table 1. *Aggregation of the 2-step paths for each direct transition taking into account the previous immediate state.*

Direct transition	Sample size	2-step transition	Sample size	Percent.
SP → Recov	223	NSP → SP → Recov	171	76.68
		SP → Recov	52	23.32
SP → NIMV	214	NSP → SP → NIMV	134	62.62
		SP → NIMV	80	37.38
SP → IMV	166	NSP → SP → IMV	92	55.42
		SP → IMV	74	44.58
SP → Death	29	NSP → SP → Death	14	48.26
		SP → Death	15	51.72
Recov → Disch	452	SP → Recov → Disch	223	49.34
		NIMV → Recov → Disch	96	21.24
		IMV → Recov → Disch	133	29.42
Recov → Death	12	SP → Recov → Death	0	0
		NIMV → Recov → Death	5	41.67
		IMV → Recov → Death	7	58.33
IMV → Death	128	SP → IMV → Death	71	57.26
		NIMV → IMV → Death	57	45.97
IMV → Recov	140	SP → IMV → Recov	95	67.86
		NIMV → IMV → Recov	45	32.14

States: NSP: No Severe Pneunomia, SP: Severe pneumonia, Recov: Severe Pneumonia recovery, NIMV: Non-Invasive Mechanical Ventilation, IMV: Invasive Mechanical Ventilation, Disch: Discharge, Death: Death

Relating the sample size of the direct transitions appearing in Table 1 with the notation introduced in Section 5, we see that the sample size of the direct transition $h \to j$ corresponds to the number of patients at-risk: $\sum_{s=2}^{T} \tilde{Y}_{hj}$, while the sample size of the 2-step transition $h \to j \to l$ does not coincide with $\sum_{s=2}^{T} \tilde{N}_{hjl}$ since we have considered all patients doing this path either in consecutive times or not.

Table 1 reveals that the proportion of patients for a given transition (e.g, IMV→ Recov) drastically differs whether the patients were before in SP (68%) or in NIMV (32%). We can also examine the transition SP→Recov, if we separate the patients between those with NSP in the admission (76.68%) and those with SP in the admission (23.32%) we also observe important differences. Similar interpretation is in place with transition SP→NIMV. This suggests that the model may not fulfill the Markov assumption and that it may be important to take the two previous states into account when calculating the transition probabilities.

6.3. Testing the Markov assumption

In order to check which 2-step transitions are not first-order Markovian, we use the Markov test described in Section 2.2 and follow Titman and Putter (2020) guidelines with respect to the time intervals $[t_0, t_{\max}]$ where the test can be conducted. Basically, the comparison is restricted to windows of time with enough individuals and to direct transitions that have an immediate previous state.

To evaluate the logrank test we compute the statistics for an equally 0.5-day spaced grid in the interval $[1, 11]$ for all the transitions except for transitions 7 (SP → Death) and 12 (NIMV → Death) in which the interval is $[1, 7]$ and transition 14 (IMV → Death) with the interval $[1, 16]$.

Table 2 summarizes the p-values of the log-rank tests obtained from 5000 wild bootstrap resamples (Lin, Wei and Ying, 1993) and considering the three possible summary statistics: weighted mean, mean, and supremum described in Section 2.2. For each transition (rows), we have carried out the test for all the possible previous states as well as the overall chi-squared test. The partial p-values are the ones corresponding to the global test. For each transition, the transition intensity compares the subjects who were previously at fixed state j (in columns) versus the ones who were not there.

Considering the overall p-values, transitions 4 (SP → Recov), 5 (SP → NIMV), 6 (SP → IMV) and 8 (Recov → Disch) show clear departures from the Markov assumption, while transition 13 (IMV → Recov) is marginally significant. Furthermore, any one of the summary tests rejects the Markovianity in transitions 4, 5 and 6. The supremum statistic would not reject Markovianity from states NSP and Recov in transition 8. Finally, the global p-value of 0.059 in transition 13 is mainly due to the non Markovianity coming from states SP and IMV.

These findings suggest that once a patient is critically ill, for instance, in states NIMV and IMV, the future clinical evolution is independent of whether he/she was diagnosed with NSP or SP when hospitalized. However, the clinical evolution to NIMV or IMV will be different for those patients initially diagnosed with NSP versus those diagnosed with

Table 2. *p-values obtained from the computation of the Markov test for each transition and each previous state. Three different summary statistics have been computed: unweighted mean (UM), weighted mean (WM) and supremum (S). In bold the transitions that are statistically significant at 0.05.*

Transitions		NSP	SP	Recov	NIMV	IMV	overall
4 (SP→ Recov)	UM	0.005	0.005				**0.0042**
	WM	0.006	0.006				
	S	0.029	0.029				
5 (SP→ NIMV)	UM	$< 10^{-16}$	$< 10^{-16}$				**0.0018**
	WM	$< 10^{-16}$	$< 10^{-16}$				
	S	0.026	0.026				
6 (SP→ IMV)	UM	0.009	0.009				**0.016**
	WM	0.002	0.002				
	S	0.042	0.042				
7 (SP→ Death)	UM	0.106	0.106				0.196
	WM	0.120	0.120				
	S	0.340	0.340				
8 (Recov→ Disch)	UM	0.007	$< 10^{-5}$	0.165	$< 10^{-5}$	$< 10^{-5}$	**$< 10^{-16}$**
	WM	0.010	$< 10^{-5}$	0.186	$< 10^{-5}$	$< 10^{-5}$	
	S	0.104	$< 10^{-5}$	0.388	$< 10^{-5}$	$< 10^{-5}$	
9 (Recov→ Death)	UM	0.652	0.298	0.145	0.495	0.143	0.357
	WM	0.644	0.273	0.144	0.464	0.151	
	S	0.656	0.313	0.353	0.639	0.309	
10 (NIMV→ Recov)	UM	0.594	0.190		0.694		0.609
	WM	0.588	0.183		0.717		
	S	0.831	0.432		0.892		
11 (NIMV→IMV)	UM	0.514	0.819		0.728		0.807
	WM	0.501	0.858		0.765		
	S	0.432	0.311		0.304		
12 (NIMV→ Death)	UM	0.348	0.218		0.649		0.437
	WM	0.378	0.253		0.619		
	S	0.338	0.342		0.719		
13 (IMV→ Recov)	UM	0.564	$< 10^{-3}$		0.514	0.005	0.059
	WM	0.531	$< 10^{-3}$		0.456	0.005	
	S	0.780	0.034		0.376	0.037	
14 (IMV→ Death)	UM	0.296	0.663		0.318	0.099	0.305
	WM	0.296	0.674		0.269	0.100	
	S	0.471	0.369		0.264	0.205	

SP when hospitalized. These results lead us to consider second-order Markov multistate models in order to study the evolution of the hospitalized COVID-19 patients during the fist wave of the pandemia.

6.4. Estimation of the transition probability matrices

We will now estimate the seven possible transition probability matrices using the estimators presented in Section 5. As we have mentioned before, we know the exact transition times, so we can easily estimate the transition probability by taking into account the number of patients who are at risk of the transition and the patients who finally have done the transition. For each row of each matrix the number of patients at risk will be different.

We present here the estimation of the matrices $\mathbf{P}_{(\mathbf{1})}$ and $\mathbf{P}_{(\mathbf{2})}$. The estimation of the rest of the matrices is similar, except for matrices $\mathbf{P}_{(\mathbf{6})}$ and $\mathbf{P}_{(\mathbf{7})}$ (6 and 7 are absorbent states) which are null matrices except for the elements $(6,6)$ and $(7,7)$ which are equal to 1.

In order to estimate the matrix $\mathbf{P}_{(\mathbf{1})}$ we start from all patients who have been hospitalized with entry in state NSP. The day after a patient has been hospitalized he/she can still be at NSP or can move to SP, Discharge or Death. So rows 1, 2, 6 and 7 are the only ones with probabilities different from 0. In order to estimate the probability cells in row 1 of $\mathbf{P}_{(\mathbf{1})}$ we consider, for each time s, all the patients who have been at least two consecutive days in NSP, that is, $\sum_{s=2}^{43} Y_{11}(s-1) = 10577$ (note here that 43 is the maximum number of days a patient has been two consecutive days in NSP). From those 10577 patients at risk, the number of patients who have stayed in NSP the next day is $\sum_{s=2}^{43} N_{111}(s) = 8919$, while $\sum_{s=2}^{43} N_{112}(s) = 257$ have transited to SP, $\sum_{s=2}^{43} N_{116}(s) = 1369$ have been discharged and, finally, $\sum_{s=2}^{43} N_{117}(s) = 32$ have died. We proceed analogously for the estimation of the probability cells in row 2 of $\mathbf{P}_{(\mathbf{1})}$ starting with those $\sum_{s=2}^{36} Y_{12}(s-1) = 411$ patients who have moved to SP from NSP the next day.

$$\widehat{\mathbf{P}}_{(1)} = \begin{pmatrix} \frac{8919}{10577} & \frac{257}{10577} & 0 & 0 & 0 & \frac{1369}{10577} & \frac{32}{10577} \\ 0 & \frac{253}{411} & \frac{3}{411} & \frac{92}{411} & \frac{62}{411} & 0 & \frac{1}{411} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

In a similar way we estimate $\mathbf{P}_{(\mathbf{2})}$. In this case we start with all patients who have been at state SP at any time. The next day these patients can still be at SP or can move to Recovery, NIMV, IMV or Death. So rows 1 and 6 are 0 because there is no direct transition from SP to NSP nor to Discharge. For row 2, the number of patients at risk, that is, the number of patients spending two consecutive times in state SP is $\sum_{s=2}^{47} Y_{22}(s-1) = 2668$. Row 3 starts with those patients who have moved from SP to Recovery, a total of $\sum_{s=2}^{47} Y_{23}(s-1) = 223$. Analogously, for rows 4 and 5, $\sum_{s=2}^{14} Y_{24}(s-1) = 214$ patients have transited immediately from SP to NIMV while $\sum_{s=2}^{23} Y_{25}(s-1) = 166$ patients moved from SP to IMV. Probability matrices $\mathbf{P}_{(\mathbf{3})}$, $\mathbf{P}_{(\mathbf{4})}$ and $\mathbf{P}_{(\mathbf{5})}$ are estimated proceeding in an analogous manner, each one starting from patients in states Recov, NIMV and IMV, respectively.

$$\widehat{\mathbf{P}}_{(\mathbf{2})} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{2307}{2668} & \frac{220}{2668} & \frac{68}{2668} & \frac{49}{2668} & 0 & \frac{24}{2668} \\ 0 & 0 & \frac{207}{223} & 0 & 0 & \frac{16}{223} & 0 \\ 0 & 0 & \frac{6}{214} & \frac{159}{214} & \frac{45}{214} & 0 & \frac{4}{214} \\ 0 & 0 & \frac{3}{166} & 0 & \frac{160}{166} & 0 & \frac{3}{166} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

6.5. Prediction via Chapman-Kolmogorov equations

The Markov test computed in Section 2.2 rejects the first-order Markov assumption for three of the four transitions from Severe Pneumonia (SP): to Non Invasive Mechanical Ventilation (NIMV), Invasive Mechanical Ventilation (IMV) and Recovery (Recov), indicating that whether or not the patient was diagnosed with Non Severe Pneumonia (NSP) marks a difference in his/her prognosis. A second-order model allows the prediction of the time to future events as a function of the diagnostic when they were hospitalized. Chapman-Kolomogorov extension in Theorem 4 is the key to the corresponding probabilities.

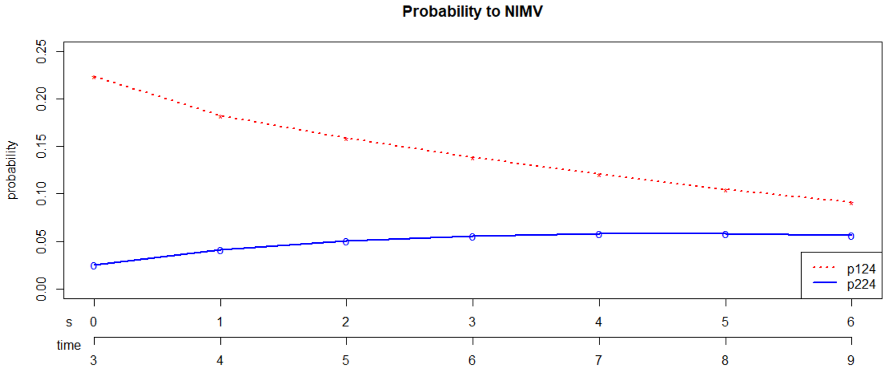

Figure 2. *Probabilities from SP to NIMV for patients who had been two consecutive days in SP (line) compared with patients who have been one day in NSP and one day in SP (dots).* $p_{224} = P(X_{3+s} = 4 | X_2 = 2, X_1 = 2)$ *vs* $p_{124} = P(X_{3+s} = 4 | X_2 = 2, X_1 = 1)$

For the transition from SP → NIMV, we will compute for $s \in \{0, \dots, 6\}$ the probabilities $p_{224} = P(X_{3+s} = 4 | X_2 = 2, X_1 = 2)$, that is, the probability to NIMV for patients who arrive at the hospital with a SP diagnosis and they still were in SP the second day. And also $p_{124} = P(X_{3+s} = 4|X_2 = 2, X_1 = 1)$ the probability to NIMV for patients with NSP at admission who had moved to SP the second day. We have plotted these probabilities in Figure 2, where we can we see how important is the initial state for the initial times. The probability of moving to NIMV of patients initially diagnosed with SP $(X_1 = 2, X_2 = 2)$ is much smaller than the probability of moving to NIMV of patients initially diagnosed with NSP $(X_1 = 1, X_2 = 2)$. These two probabilities close the gap as days go by.

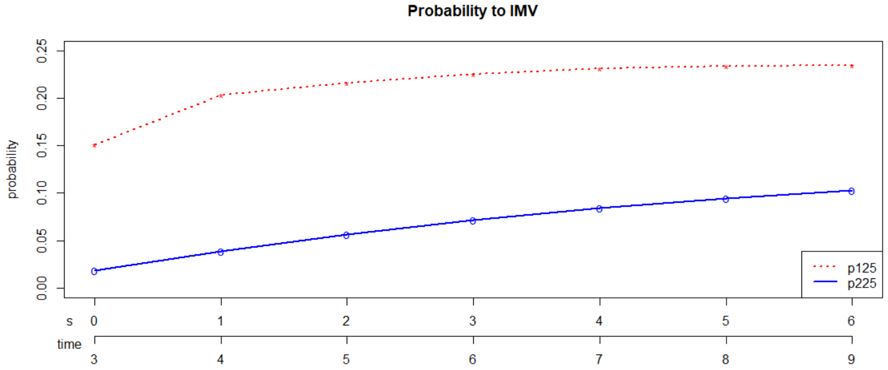

Figure 3. *Probabilities from SP to IMV for patients who had been two consecutive days in SP (line) compared with patients who have been one day in NSP and one day in SP (dots).* $p_{225} = P(X_{3+s} = 5|X_2 = 2, X_1 = 2)$ *vs* $p_{125} = P(X_{3+s} = 5|X_2 = 2, X_1 = 1)$.

The same type of plot is depicted in Figure 3 to study the transition SP → IMV. In this case the patients are also splitted based on their initial state: NSP or SP. As in Figure 2, the probability of moving to IMV of patients initially diagnosed with SP $(X_1 = 2, X_2 = 2)$ is much smaller than the probability of moving to IMV of patients initially diagnosed with NSP $(X_1 = 1, X_2 = 2)$. However, both probabilities increase over the time and their difference is kept along the next days. This reveals the different prognosis for needing respiratory mechanical ventilation (IMV) among those patients initially diagnosed with NSP versus being diagnosed with SP.

7. Discussion

In this paper we have introduced a second-order Markov multistate model and we have developed an extension of the Chapman-Kolmogorov equations to compute r-step transition probabilities. We have used the DIVINE COVID-19 data to estimate the transition probabilities and to predict probabilities to NIMV and IMV in terms of the states where a patient was during the first 2 days of his/her hospitalization.

As we briefly mention in the introduction, a second-order Markov model could had been transformed into a first-order Markov model by redefining the state space. This would be possible creating extra states formed by direct 1-step transitions. For instance, in the DIVINE data case instead of one unique Death state we could have defined 3 new states formed by those patients arriving to Death from NIMV, IMV or Recovery. The advantage of these new states is clear because we would be able to apply all the knowledge on first-order Markov models. However, the number of states and transitions of the new model will increase substantially and the interpretation will become cumbersome. Furthermore, since the number of parameters to estimate will increase and, so does the needed sample sample size to estimate all of them, the second-order Markov approach is preferable.

It should be mentioned that Chapman-Kolmogorov extension is based on a discretization of the time scale and is only computed conditionally to two-consecutive times. But, we prove that one-step transition probabilities and one-step second-order transition probabilities together with the extended Chapman-Kolmogorov equations are enough to compute the transition probabilities if the previous two times are not consecutives.

In this paper we sketch two different ways to estimate the transition probabilities. The first one using the Bernoulli probabilities, which is the one used to compute the transition probabilities in the COVID illustration example and the second using conditional probabilities. Since the data from the DIVINE project was collected one year after the end of the first wave, we have complete registries and, for this reason, we have so far only developed both methods for complete (uncensored) data. Nevertheless, it is indeed relevant to extend these estimators to account for right-censored data. The second method of estimation presented in Subsection 5.2 gives a clue of how we could proceed to account for right-censored data. This estimator, an average of the ratios, for each time, of those subjects doing an specific transition among the number of subjects at risk, has an analogy to the Nelson-Aalen estimator for the cumulative hazard function. For a thorough statistical analysis, the derivation of the variance of these estimators as well as of their asymptotic distribution is needed. Furthermore, estimators for the state occupation probabilities and for the transition intensities for complete and right-censored data are as well a topic of interest. All these ideas remain open for our future research.

Acknowledgements

This research has been funded by the Ministerio de Ciencia e Innovación (Spain) [PID 2019-104830RB-I00/ DOI(AEI): 10.13039/501100011033] and by Generalitat de Catalunya through the projects 2020PANDE00148 and 01421 SGR-Cat 2021. We are indebted to our colleagues in the DIVINE group for their clever contributions and dedicated time.

References

Chakladar, S., R. Liao, W. Landau, M. Gamalo, and Y. Wang (2022). Discrete Time Multistate Model With Regime Switching for Modeling COVID-19 Disease Progression and Clinical Outcomes. *Statistics in Biopharmaceutical Research*, 14, 52–66.

Ching, W. K., Fung, E. S. and Ng, M. K. (2003). A higher-order Markov model for the Newsboy's problem. *Journal of the Operational Research Society*, 54, 291–298.

de Wreede, L., 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, 1–30.

Garmendia, L., Cortés, J. and Gómez Melis, G. (2023). MSMpred: Interactive modelling and prediction of individual evolution via multistate models. *BMC Medical research methodology*, 23.

Hougaard, P. (1999). Multi-state models: a review. *Lifetime Data Anal.*, 5, 239–264.

Kay, R. (1986). A Markov Model for Analysing Cancer Markers and Disease States in Survival Studies. *Biometrics*, 42, 855–865.

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

Logan, J. A. (1981). A structural model of the higher-order Markov process incorporating reversion effects. *The Journal of Mathematical Sociology*, 8, 75–89.

Mody, A., Lyons, P. G., Vazquez Guillamet, C., Michelson, A., Yu, S., Namwase, A. S., Sinha, P., Powderly, W. G., Woeltje, K. and Geng, E. H. (2020). The Clinical Course of Coronavirus Disease 2019 in a US Hospital System: A Multistate Analysis. *American Journal of Epidemiology*, 190, 539–552.

Pallarès, N., Tebé, C., Abelenda-Alonso, G., Rombauts, A., Oriol, I., Simonetti, A. F., Rodríguez-Molinero, A., Izquierdo, E., Díaz-Brito, V., Molist, G., Gómez Melis, G., Carratalà, J., Videla, S. and study groups, M. (2023). Characteristics and Outcomes by Ceiling of Care of Subjects Hospitalized with COVID-19 During Four Waves of the Pandemic in a Metropolitan Area: A Multicenter Cohort Study. *Infectious diseases and therapy*, 12, 273–289.

Piulachs, X., Langhor, K., Besalú, M., Pallarès, N., Carratalà, J., Tebé, C. and Gómez Melis, G. (2023). Semi-Markov multistate approaches for multicohort event history data. *(submitted)*.

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

Shamshad, A., Bawadi, M., Wan Hussin, W., Majid, T. and Sanusi, S. (2005). First and second order markov chain models for synthetic generation of wind speed time series. *Energy*, 30, 693–708.

Shorrocks, A. F. (1976). Income Mobility and the Markov Assumption. *The Economic Journal*, 86, 566–578.

Titman, A. C. and Putter, H. (2020). General tests of the Markov property in multi-state models. *Biostatistics*, 23, 380–396.

Tong, H. (1975). Determination of the order of a Markov chain by Akaike's information criterion. *Journal of Applied Probability*, 12, 488–497.