BACHELOR'S DEGREE THESIS

Double degree in Statistics and Economics

Title: Time-homogeneous Markov Multi-state models

Author: Aitor Moruno Cuenca

Advisor: Mireia Besalú Mayol
Co-Advisor: José Ramón García Sanchís

Advisor's department: Genetics, Microbiology and statistics: section of statistics.
Co-Advisor's Department: Econometrics, statistics and applied economics.

Academic year: 2021-2022

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

University of Barcelona
Polytechnic University of Catalonia

Bachelor's thesis
Double degree in Statistics and Economics

Time-homogeneous Markov Multi-state models

By

Aitor Moruno Cuenca

Advisor:
Mireia Besalú Mayol

Co-Advisor:
José Ramón García Sanchís

28 june of 2022
Academic year 2021-2022

Agraïments

Primer, vull donar les gràcies a Mireia per haver accedit a dirigirme aquest treball però, sobretot, per la seva predisposició i implicació, fins i tot quan li enviaba correus infinits. He gaudit i après molt fent aquest treball, i això és, en part, gràcies a ella. També voldria donar les gràcies a en José Ramón per haver accedit a co-tutoritzar aquest treball i pels seus comentaris sempre que li he requerit ajuda.

A Berta i a Emma, que sempre heu confiat en mi i m'heu donat el suport que he necessitat en els moments més difícils al llarg d'aquests anys, us vull agrair que formeu part de la meva vida. Tanmateix, gràcies Gaizka i Alba pel vostre suport durant aquest any, a més dels teus valuosos comentaris respecte el treball Alba.

Als companys i amics del grau que, durant aquests 5 anys han estat imprescindibles, hagués estat tot molt més difícil sense vosaltres. En especial us dono les gràcies a Arnau, Cesc, Carlos i Ferran.

Abstract

Bachelor's thesis in Statistics and Economics

Time-homogeneous Markov Multi-state models

by Aitor Moruno Cuenca

A standard survival analysis model can be represented as a multi-state model with two states (alive and dead) and one transition between them alive to dead. Multi-state models generalize survival analysis models by introducing multiple states and transitions, thus providing information on the dynamics of interest. These models are very flexible, they can model almost any kind of situation with longitudinal data, even introducing multiple censoring patterns and covariates, they find applications on a wide variety of fields: insurance, medical statistics, politics, epidemiology. In this thesis we focus on a multi-state model driven by a time-homogeneous Markov process, this assumption leads to well-known survival analysis tools like Cox proportional-hazards model. A detailed introduction to these models is presented; from its construction with counting and Markov processess, to an application to trait-based vaccination strategies against tuberculosis on wild animals. With respect to the theory of multi-state models, contributions are made by providing non-existent proofs from the references used in this thesis. Further, the application is based on Patterson SJ, et.al (2022) [14] but from a multi-state methodological point of view.

Keywords: *Multi-state models, Markov process, survival analysis, medical statistics, transition intensity functions, censoring, counting process, product integration*

AMS Classification (2010): 60J27 Continuous-time Markov processes on discrete state spaces

60J28 Applications of continuous-time Markov processes on discrete state spaces

62M05 Markov processes: estimation; hidden Markov models

62N01 Censored data models

62N02 Estimation in survival analysis and censored data

Resum

Treball fi de grau en Estadística i Economia

Time-homogeneous Markov Multi-state models

per Aitor Moruno Cuenca

Un model estàndard d'anàlisi de supervivència es pot representar com un model multi-estat amb dos estats (viu i mort) i una transició entre aquests estats: viu a mort. Els models multi-estat, generalitzen els models de supervivència introduint múltiples estats i transicions, així doncs, permeten extreure informació sobre la dinàmica del procés d'interés. Aquests models permeten modelitzar pràcticament qualsevol situació amb dades longitudinals, fins i tot introduint múltiples patrons de censura i covariables, i troben aplicacions en camps tan diversos com: assegurances, estadística mèdica, política, epidemiologia. En aquest treball ens centrem en un model multi-estat dirigit per un procés de Markov homogeni a temps continu, la selecció d'aquest tipus de procés permet emprar eines ja conegudes d'anàlisi de supervivència com el model de riscos proporcionals de Cox. Es presenta una introducció teòrica i completa a aquests models; des de la seva construcció amb processos comptadors i de Markov, fins una aplicació a l'elecció de l'estratègia òptima de vacunació per a la tuberculosi en animals salvatges. Pel que fa a l'estudi teòric d'aquests models, es detallen i es fan aportacions de demostracions no existents en la literatura emprada en aquest treball, ja sigui per omissió o per manca de detall, de manera que tots els conceptes tractats quedaran reflectits en aquestes. D'altra banda, l'aplicació està basada en Patterson SJ, et.al (2022) [14] però des d'una perspectiva metodològica dels models multi-estat.

Paraules clau: *Models multi-estat, processos de Markov, anàlisi de supervivència, estadística mèdica, funcions de riscos de transició, censura, processos comptadors, integrals producte.*

Contents

List of Figures

List of Tables

Introduction

When working with some acute diseases, such as advanced cancer or fatal diseases, there is often a single dominating endpoint, in such case Kaplan-Meier curves and Cox models are sufficient and efficient tools for the analysis, but when multiple important endpoints arise or the event in question may occur more than once for an individual, the former tools become obsolete. The need for a proper statistical metholodogy, in such cases, takes its form in multi-state models.

Longitudinal data is frequent in chronic disease studies where patients are observed over time and discrete states, based on clinical significance, are recorded. A change on the intercourse of the disease is called a transition. Multi-state models (MSM) have become a powerful tool in the analysis of longitudinal event history data, these models extend survival models. In survival models, there are just two possible states at any time point considered: alive and dead, and one transitions alive to dead. MSM allow to extract information on the dynamics of any process with a finite number of states and transitions, which is a framework quite suitable for fields like insurance and biostatistics, although in this thesis we mainly focus on biostatistical applications.

One of the primary goals of this thesis is to study the theoretical foundation of such models, and implement them to an existent dataset through `msm` package from `R` software. Since MSM are really broad, we mostly limit this thesis to the study of MSM driven by a time-homogeneous Markov process, this assumption also aims to give a smoother introduction to MSM otherwise it could get mathematically very challenging. Furthermore, we aim to detail theoretical results skipped by the primary references, and contribute to non-existent proofs on them. Take into account that all proofs and problem resolutions that are taken over from another work include a reference to the original source in the proof, otherwise no reference will be given. This thesis is not an introduction to more advanced topics on multi-state models (MSM) such as: time-varying covariates, design issues, censoring patterns beyond (at most) right-censoring, goodness of fit tests for MSM, non-homogeneous Markov MSM, and so on, as the primary objective is to understand the most basic foundations of these models. Once the theoretical framework is set, we will apply the MSM methodology to an existent dataset extracted from [14], simplifications of the models fitted will be made when deemed necessary if the right methodology goes beyond the scope of the thesis.

This work is divided into four chapters; the first three chapters are the theoretical framework on which MSM are set, chapter four is an application of MSM to trait-based vaccination strategies in individual meerkats. In chapter 1, definitions and results needed to develop MSM theory are studied, concepts that will be reviewed in this chapter are: product integral, continuous time Markov process, properties of random variables, censoring patterns, stochastic processes, and so on. In the next chapter, we motivate the use of MSM through some real clinical studies that have carried out their analysis with MSM methdology, after that we start building from scratch a MSM by using three stochastic processes: two counting processes and a continuous time Markov process. To end with the theoretical part, in chapter 3 we detail the steps to derive the likelihood of a general MSM with, at most a right-censoring pattern, and sketchily present the estimation procedure for a MSM driven by a homogeneous Markov process. To end, in chapter 4 we apply MSM methodology to a real dataset through `R` software, the full process of a statistical analysis is presented: dataset structures for MSM, design of the study, primary objectives of the analysis, data cleaning, descriptive analysis, proposed metholodogy and results. The dataset is extracted from a peer-reviewed article, but the statistical methodology proposed is completely different from the one treated in the article, thus no analysis used in the article has been utilized.

Chapter 1

Mathematical and statistical preliminaries

1.1 Survival analysis

Survival analysis studies the time until an event $\mathcal{E}$ of interest occurs, such an event $\mathcal{E}$ could be: time until death, the diagnosis of a tumor, machine failure, and so on. In this section, we will revise some useful results and definitions in Survival analysis theory that will be used in multi-state models.

1.1.1 Definitions and results

Definition 1.1.1. *(Survival function) The survival function of a random variable T is defined as the probability that T is greater than a certain value t, that is the probability that the event occurs posterior to t: $S(t) = \mathbb{P}(T > t) = 1 - F(t)$.*

Definition 1.1.2. *(Hazard function) For an absolutely continuous random variable T, the hazard function $\lambda(t)$ is defined as*

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T \leq T + \Delta t)}{\Delta t}$$

If T is a discrete random variable taking values $0 = t_1 < t_2 < \dots$ the hazard function is the probability that the event occurs in t_{j+1} given that it has not occurred in t_j: $\lambda(t_j) = \mathbb{P}(T = t_{j+1} | T > t_j)$.

Definition 1.1.3. *(Cumulative hazard function) For an absolutely continuous random variable T the cumulative hazard function is $\Lambda(t) = \int_0^t \lambda(s)ds$. If T is a discrete random variable with values $0 = t_1 < t_2 < \dots$ the hazard function is defined as $\Lambda(t) = \sum_{j:t_j \leq t} \lambda(t_j)$*

Proposition 1.1.1. *Let T be a positive random variable with survival function $S(t)$, such that for all $k > 0$ the k-th order moment exists and its finite $\mathbb{E}[T^k] = \mu_k < \infty$ and $\lim_{t \to \infty} S(t^{1/k}) = 0$. Then the following property holds*

$$\mathbb{E}[T^k] = \int_0^\infty S(u^{1/k})du$$

Proof. If T is an absolutely continuous random variable, write $T^k = \int_0^\infty I_{\{T^k > u\}}(u)du$ then its expectation is given by

$$\begin{aligned}\mathbb{E}[T^k] = \mathbb{E}\left[\int_0^\infty I_{\{T^k>u\}}(u)du\right] &= \int_0^\infty \left(\int_0^\infty I_{\{t^k>u\}}(u)du\right) f(t)dt \overset{(a)}{=} \int_0^\infty \int_0^\infty I_{\{t^k>u\}}(u)f(t)dtdu \\ &= \int_0^\infty \int_{u^{1/k}}^\infty f(t)dtdu = \int_0^\infty \mathbb{P}(T > u^{1/k})du = \int_0^\infty S(u^{1/k})du\end{aligned}$$

Where I have used Fubini's Theorem in (a). To justify step (a) it suffices to show that $\int_0^\infty S(u^{1/k})du$ is finite.

$$\int_0^\infty S(u^{1/k})du = \int_0^\infty (1-F(u^{1/k}))du \overset{(b)}{=} (1-F(u^{1/k}))u\Big|_{u=0}^{u\to\infty} - \int_0^\infty -\frac{uf(u^{1/k})u^{\frac{1}{k}-1}}{k}du$$
$$\overset{(c)}{=} \frac{1}{k}\int_0^\infty u^{1/k}f(u^{1/k})du \overset{(d)}{=} \mu_k < \infty$$

(b) Integration by parts using $t = 1-F(u^{1/k})$, $dt = -\frac{dF(u^{1/k})}{du} = -\frac{d}{du}\int_0^{u^{1/k}} f(t)dt = -f(u^{1/k})\frac{d}{du}u^{1/k} = -\frac{f(u^{1/k})u^{\frac{1}{k}-1}}{k}du$, and $dv = du$ $v = u$

(c) The limit part

$$0 \leq \lim_{u\to\infty} u(1-F(u^{1/k})) = \lim_{u\to\infty} uS(u^{1/k}) = \lim_{u\to\infty} u\int_{u^{1/k}}^\infty f(t)dt$$
$$= \lim_{u\to\infty}\int_{u^{1/k}}^\infty uf(t)dt \leq \lim_{u\to\infty}\int_{u^{1/k}}^\infty t^k f(t)dt = 0$$

Inequality follows from $u \leq t^k$ $\forall k > 0$, monotonicity of Riemann integral and because limits preserve inequalities.

(d) By assumption the moment of k-th order exists and its finite then

$$\mu_k = \int_0^\infty t^k f(t)dt = \frac{1}{k}\int_0^\infty tf(t)(kt^{k-1}dt) = \frac{1}{k}\int_0^\infty w^{1/k}f(w^{1/k})dw$$

I have applied a change of variables $w = t^k, dw = kt^{k-1}$.

If T is a discrete random variable taking values $0 = t_1 < t_2 < \ldots$, then its k-th order moment is given by:

$$\mathbb{E}[T^k] = \sum_{j=1}^\infty t_j^k \mathbb{P}(T = t_j) = \sum_{j=1}^\infty u_j \mathbb{P}(T = u_j^{1/k}) = \sum_{j=1}^\infty u_j(S(u_j^{1/k}) - S(u_{j+1}^{1/k}))$$
$$= \sum_{j=1}^\infty u_j S(u_j^{1/k}) - \sum_{j=1}^\infty u_j S(u_{j+1}^{1/k}) = \sum_{j=2}^\infty u_j S(u_j^{1/k}) - \sum_{j=1}^\infty u_j S(u_{j+1}^{1/k})$$
$$= \sum_{j=2}^\infty u_j S(u_j^{1/k}) - \sum_{j=2}^\infty u_{j-1} S(u_j^{1/k}) = \sum_{j=2}^\infty (u_j - u_{j-1})S(u_j^{1/k}) = \int_0^\infty S(u^{1/k})du$$

In the last step, the Riemann sum converges as $\lim_{u\to\infty} S(u^{1/k}) = 0$ and $\mathbb{E}[T^k]$ exists and its finite. □

Proposition 1.1.2. *If T is an absolutely continuous random variable, the survival function can be expressed as $S(t) = \exp\{-\Lambda(t)\}$.*

Proof. By the definition of hazard function $\lambda(t) = -\frac{d\log(S(t))}{dt} = -\frac{S'(t)}{S(t)} \iff S'(t) = -\lambda(t)S(t)$, which is a first order linear homogeneous differential equation with initial value $(t_0, S(t_0)) = (0, 1)$, hence

$$S(t) = S(t_0)\exp\left\{-\int_{t_0}^t \lambda(u)du\right\} = \exp\left\{-\int_0^t \lambda(u)du\right\} = \exp\left\{-\Lambda(t)\right\}$$

□

Definition 1.1.4. *(Cox proportional hazards model) Let $\mathbf{X}$ be a $(p\times 1)$-dimensional vector of covariates and β a $(p\times 1)$-dimensional vector of parameters, a Cox proportional hazards model is a multivariate regression over time between hazard functions and covariants as follows*

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\exp\{\beta'\mathbf{X}\} \tag{1.1}$$

Where $\lambda_0(t) = \lambda(t|\mathbf{X} = 0)$ is known as the baseline hazard, which is the hazard of an individual with profile $\mathbf{X} = 0$.

1.1.2 Right-censored data

Censoring is a statistical condition in which the value of a measurement or observation is only partially known. There are many types of censoring patterns, but only right-censoring will be discussed throughout the whole study, as it is the easiest and most common one, for extra information about other censoring patterns (left-censored or interval-censored data) one can refer to [13]. Throughout the coming subsections we will assume:

- T as the survival time, a non-negative variable, that can either be continuous or discrete.
- For each individual i we define a survival time T_i and a censoring time C_i. We will consider the case that censoring is independent (non-informative), so T_i is independent of C_i.
- $T_1, T_2, \ldots, T_n$ (n number of individuals) are independent and identically distributed with unknown distribution function F.
- We will observe a value Y_i defined as $Y_i = \min\{T_i, C_i\}$ and an indicator of censoring δ_i as $\delta_i = I_{\{T_i \leq C_i\}}$, which takes value 1 if Y_i is the survival time and 0 if it is the time to censoring.
- The sample will be usually presented by the pair $\{(Y_i, \delta_i)\}_{i=1}^n$

Now, we will present the different types of right censoring.

Type I censored data

We say that an event has *censoring of type I* when true survival time is only observed if it happens before a prior specified time C^A, we will only know the true survival time of an individual when $T_i \leq C^A$, when $T_i > C^A$ the subject has survived and the observation is censored. Formally, the sample is defined by a pair of random variables (Y_i, δ_i) where $Y_i = \min\{T_i, C^A\}$ and $\delta_i = I_{\{T_i \leq C^A\}}$, δ_i is known as *censoring indicator*.

Another subtype of censoring of type I is *progressive censoring of type I*, when C^A differs from each group of individuals. A set of administrative censoring times is defined for each different group, $C_1, \ldots, C_m$ where $m \leq n$, where n is the number of individuals grouped in m groups. A common study where this censoring applies is in animal studies, where different sacrifice times are set for each animal group so as to assess the existence of a non-palpable and clinically hidden tumour. Formally, define a finite number of censoring times $C_1, \ldots, C_m$ and subjects are classified in m different groups in such a way that every group has a different censoring time, the subjects $\{1, \ldots, n_1\}$ form the group 1 with censoring time C_1, subjects $\{n_1 + 1, \ldots, n_2\}$ form the group 2 with censoring time C_2, and so on until the group m formed by subjects $\{n_{m-1} + 1, \ldots, n_m\}$ with censoring time C_m. We observe a sample $\{(Y_i, \delta_i)\}_{i=1}^n$ where for each $j = 1, \ldots, m$ and $n_{j-1} + 1 \leq i \leq n_j$, $Y_i = \min\{T_i, C_j\}$ and $\delta_i = I_{\{T_i \leq C_j\}}$.

Generalized when patients are recruited sequentially in the study, but there is a common time for all of them at which study will end denoted by C^A. Usually, we re-scale the survival time, to consider the same entrance time for all individuals. This censoring is common in clinical trials where patients are recruited sequentially but they all face the same end of study date. Formally, denote $\mathcal{O}_i$ the entrance time of the individual i with censoring time $C_i = C^A - \mathcal{O}_i$, the sample is then $\{(Y_i, \delta_i)\}_{i=1}^n$ where $Y_i = \min\{Y_i, \delta_i\}$ and $\delta_i = I_{\{T_i \leq C_i\}}$.

Type II censored data

Type II occurs when the design of the study finishes when the event has been observed r times, previously defined, the number of subjects in the study is less than the specified number of times the event is observed: $r < n$. This is kind of censoring is common in an engineering framework, where observing failure is costly and a maximum amount of observed failure must be imposed to reduce costs. The time until event is observed is random, but the number of observed and censored events is not random as it is a fixed number r and $n - r$, respectively. Formally, consider $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$ the ordered survival times, the random censoring time $C = T_{(r)}$, it is the same for all individuals, we observe a sample $\{(Y_i, \delta_i)\}_{i=1}^n$ where $Y_i = \min\{T_i, T_{(r)}\}$ and $\delta_i = I_{\{T_i \leq T_{(r)}\}}$.

Type III censored data

Type III is also known as *random censoring*. This censoring arises when we are interested in a marginal distribution of an event but subjects are lost to follow-up due to other events not considered, we name those events as competitive events. This is common in a clinical trial setting, patients can observe death due to many other causes rather than illness itself: car crash, serious adverse event, other illnesses, and so on. The event of interest is not observed if subjects experience a competetitive event, thus there exists a right censoring. Formally, as stated before, $T_1, \ldots, T_n$ is independent and identically distributed with unknown distribution F, furthermore $C_1, \ldots, C_n$ are independent and identically distributed with unknown distribution G. Also, T_i and C_i are independent $\forall i = 1, \ldots, n$. We observe a sample $\{(Y_i, \delta_i)\}_{i=1}^n$ with $Y_i = \min\{T_i, C_i\}$ and $\delta_i = I_{\{T_i \leq C_i\}}$.

1.2 Stochastic processes

Some elementary results and definitions of Stochastic Processes, that are used throughout the coming chapters, will be introduced.

Definition 1.2.1. *(Filtration) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(I, \leq)$ be a totally ordered set. For every $k \in I$ let $\mathcal{F}_k$ be a sub-σ-algebra of $\mathcal{F}$. Then*

$$\mathbb{F} := (\mathcal{F}_k)_{k \in I}$$

is called a filtration, if for all $k \leq l$ then $\mathcal{F}_k \subseteq \mathcal{F}_l$

Definition 1.2.2. *(Stochastic process) A stochastic process X is a collection of random variables $\{X(t)\}_{t \in I}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, indexed by some set I (index set), and a measurable space $(S, \mathcal{S})$, where S is known as state space, where $X(t) : \Omega \to S$*

We will introduce an example of a filtration that will be extensively used in multi-state models. Let $(N(t))_{t \geq 0}$ be a stochastic process on the probability space $(\mathbb{N}, \mathcal{P}(\mathbb{N}), \mathbb{P})$ and $\mathbb{H} := (\mathcal{H}_k)_{k \in \mathbb{N}}$ with $\mathcal{H}_k := \sigma(N(s) : 0 \leq s \leq k)$, we will prove that $\mathbb{H}$ defines a filtration. To show this family of sub-σ-algebras defines a filtration it suffices to prove that $\mathcal{H}_k \subseteq \mathcal{H}_l$ whenever $k \leq l$ for $k, l \in \mathbb{N}$

$$\begin{aligned}
\mathcal{H}_l := \sigma(N(s) : 0 \leq s \leq l) &= \sigma(N(s) : \{0 \leq s \leq k\} \cup \{k \leq s \leq l\}) \\
&\supseteq \sigma(N(s) : 0 \leq s \leq k) \cup \sigma(N(s) : k \leq s \leq l) \\
&= \mathcal{H}_k \cup \{\mathcal{H}_l \backslash \mathcal{H}_k\} \supseteq \mathcal{H}_k
\end{aligned}$$

For the inclusion I have used that for any collection of sets $\{A_t\}_{t \in I}$ it holds $\bigcup_{t \in I} \mathcal{P}(A_t) \subseteq \mathcal{P}(\bigcup_{t \in I} A_t)$, indeed pick any $a \in \bigcup_{t \in I} \mathcal{P}(A_t)$ then $\exists t_0 \in I$ such that $a \in \mathcal{P}(A_{t_0})$ as a is an element of the power set of A_{t_0} necessarily $a \in A_{t_0} \subseteq \bigcup_{t \in I} A_t$, which yields to $a \in \mathcal{P}(\bigcup_{t \in I} A_t)$ that proves the desired inclusion.

Definition 1.2.3. *(a.s right-continuous) A stochastic process $\{X(t)\}_{t \in I}$ is a.s right-continuous if $\forall t \in I$ $\mathbb{P}(\{\omega \in \Omega : \lim_{\varepsilon \to 0} |X(t, \omega) - X(t + \varepsilon, \omega)| = 0\}) = 1$*

1.2.1 Counting processes

Survival analysis theory can be fully expressed using counting process notation. In this section definitions and results used throughout the coming chapters will be presented.

Definition 1.2.4. *(Counting process) A stochastic process $N := \{N(t)\}_{t \geq 0}$ is called a counting process if*

- $N(0) = 0$
- $N(t) < \infty, \forall t \geq 0$
- *$N(t)$ is an a.s right-continuous step function with jumps of size 1*
- *for $0 \leq s < t$, $N(t) - N(s)$ gives the number of events in the interval $(s, t]$*

Proposition 1.2.1. *(Integral with respect to counting process) Let $K(t)$ be a random function of time, then the integral of $K(t)$ with respect to the counting process $N(t)$*

$$\int_0^t K(s)dN(s) = \sum_{\{1\leq i\leq n: dN(t_i)=1\}} K(t_i)$$

Where t_i for $i = 1, \ldots, n$ are jump times at which $dN(t_i) = 1$.

1.2.2 Markov processes

This section will be primarily devoted to continuous time Markov chains, as these are the most studied processes in multi-state models. The starting point is the definitions of: Markov property, discrete time Markov chain, continuous time Markov chain, after those, main results of continuous time Markov chains will be studied.

Definition 1.2.5. *(Markov property). A stochastic process $\{X(t)\}_{t\in I}$, is said to possess the Markov property if, for each $A \in \mathcal{S}$ and each $s, t \in I$*

$$\mathbb{P}(X(t+s) \in A|\mathcal{F}_s) = \mathbb{P}(X(t+s) \in A|X(s))$$

Definition 1.2.6. *(Discrete time Markov chain). A stochastic process $\{X(t)\}_{t\in I}$ is a discrete time Markov chain if:*

- *Satifies the Markov property*
- *Discrete State Space, usually $S \subseteq \mathbb{N}$ or $S \subseteq \mathbb{Z}$*
- *Countable index set I, usually $I = \mathbb{N}$ or $I = \mathbb{Z}$*

Definition 1.2.7. *(Continuous time Markov chain). A stochastic process $\{X(t)\}_{t\in I}$ is a continuous time Markov chain (CTMC) if:*

- *Satifies the Markov property*
- *Discrete State Space S, usually $S \subseteq \mathbb{N}$ or $S \subseteq \mathbb{Z}$*
- *Non-countable index set I, usually $I \subseteq \mathbb{R}$*

Definition 1.2.8. *(Time-homogeneity). A CTMC is time-homogeneous if for any $s \leq t$ and $\forall i, j \in S$*

$$\mathbb{P}(X(t) = i|X(s) = j) = \mathbb{P}(X(t-s) = i|X(0) = j)$$

Intuitively, whenever the process enters state i, the probabilities from that point is the same as if the process started in state i at time 0, hence time does not affect probabilities between states. For non-homogeneous CTMC probabilities depend on both time and current state the process is in (i.e probabilities evolve over time).

Definition 1.2.9. *(Transition probability). The transition probabilities for a homogeneous CTMC is defined for every $t > 0$ and is denoted as*

$$p_{ij}(t) = \mathbb{P}(X(t) = j|X(0) = i)$$

From now on, we will devote the coming results and definitions to homogeneous CTMC if not stated otherwise. The following proposition comes up with the idea that, for a homogenous CTMC, it suffices to know the transition probabilities $p_{ij}(t)$ for all $t < t_0$, for a fixed $t_0 > 0$, to deduce the rest of the transition probabilities $p_{ij}(t)$ for all t.

Proposition 1.2.2. *(Chapman-Kolmogorov equation). Let $\{X(t) : t \geq 0\}$ be a CTMC with state space S and transition probabilities $(p_{ij}(t))_{i,j\in S}$. Then for any $s, t \geq 0$,*

$$\sum_{k\in S} p_{kj}(t)p_{ik}(s) = p_{ij}(s+t)$$

Proof. I follow proof given by [12] in Chapter 2. Using conditional probability and Markov property

$$\begin{aligned} p_{ij}(s+t) = \mathbb{P}(X(s+t)=j|X(0)=i) &= \sum_{k\in S} \mathbb{P}(X(s+t)=j, X(s)=k|X(0)=i) \\ &= \sum_{k\in S} \mathbb{P}(X(s+t)=j|X(s)=k, X(0)=i)\mathbb{P}(X(s)=k|X(0)=i) \\ &= \sum_{k\in S} p_{kj}(t)p_{ik}(s) \end{aligned}$$

□

Definition 1.2.10. *(Transition intensity). The transition intensity from state $i \to j$ is defined as the limit:*

$$q_{ij} = \lim_{h\to 0} \frac{p_{ij}(h)}{h}$$

Assuming the limit $lim_{h\to 0}\frac{p_{ij}(h)}{h} < \infty$.

This last definition gives the idea on how to construct the probabilities $p_{ij}(t)$ for all time t, which is defined as its derivative with respect to time.

Definition 1.2.11. *(Stochastic matrix). A matrix $P = \{p_{ij}\}_{i,j\in S}$ is stochastic if it satisfies:*

- $0 \leq p_{ij} < \infty, \forall i, j \in S$
- $\sum_{j\in I} p_{ij} = 1, \forall i \in S$

An example of a Stochastic matrix is a matrix that contains the transition probabilities of a CTMC so as $P(t) = \{p_{ij}(t)\}_{i,j\in S}$ is a stochastic matrix for all t.

Definition 1.2.12. *(Transition intensity matrix). Let S be a countable set. A transition intensity matrix on S is a matrix $Q = \{q_{ij}\}_{i,j\in S}$ that satisfies:*

- $0 \leq -q_{ii} < \infty, \forall i \in S$
- $q_{ij} \geq 0, \forall i, j \in I, i \neq j$
- $\sum_{j\in I} q_{ij} = 0, \forall i \in S$

An example of a transition intensity matrix is a matrix $Q = (q_{ij})_{i,j\in S}$, note that this matrix does not depend on time contrary to $P(t) = \{p_{ij}(t)\}_{i,j\in S}$. Each row of Q sums up to 0, as the sum of every row is finite $q_i = \sum_{j\neq i} q_{ij} < \infty$ and the diagonal element is $q_{ii} = -q_i$, then the sum of the row is equal to 0.

In fact, transition intensity matrices and stochastic matrices are related to each other by the following theorem

Theorem 1.2.1. *A matrix Q on a finite set S is a transition intensity matrix if and only if $P(t) = e^{tQ}$ is a stochastic matrix for all $t \geq 0$.*

Theorem 1.2.2. *Let Q be a matrix on a finite set S, define $P(t) = e^{tQ}$. Then, the process $(P(t))_{t\geq 0}$ satisfies:*

- *$(P(t))_{t\geq 0}$ is the unique solution for the Forward-Kolmogorov differential equation (FKE)*

$$\begin{cases} \frac{d}{dt}P(t) = P(t)Q \\ P(0) = I \end{cases} \tag{1.2}$$

- *$(P(t))_{t\geq 0}$ is the unique solution for the Backward-Kolmogorov differential equation (KBE)*

$$\begin{cases} \frac{d}{dt}P(t) = QP(t) \\ P(0) = I \end{cases} \tag{1.3}$$

Proof. I follow proof given by [12] in Chapter 2, I detail on the radius of convergence and explicit computation of the derivative of the power series . We first prove existence FKE. By hypothesis $P(t) = e^{tQ}$, using definition of the exponential of a matrix

$$P(t) = \sum_{k=0}^{\infty} \frac{(tQ)^k}{k!} = \sum_{k=0}^{\infty} a_k (tQ)^k \tag{1.4}$$

The radius of convergence of (1.4) is $\mathbb{R}$:

$$R = \lim_k \frac{|a_k|}{|a_{k+1}|} = \lim_k \frac{(k+1)!}{k!} = +\infty$$

And thus $P(t) \in C^\infty(\mathbb{R})$, which implies $P(t)$ is term-by-term differentiable for $\forall t \geq 0$.

$$P'(t) = \frac{d}{dt} \sum_{k=0}^{\infty} \frac{(tQ)^k}{k!} = \sum_{k=0}^{\infty} \frac{d}{dt} \frac{(tQ)^k}{k!} = \sum_{k=1}^{\infty} \frac{kt^{k-1}Q^k}{k!} = \sum_{k=1}^{\infty} \frac{k(tQ)^{k-1}Q}{k(k-1)!} \overset{(a)}{=} \left(\sum_{i=0}^{\infty} \frac{(tQ)^i}{i!} \right) Q = P(t)Q$$

(a) Change of index $i = k - 1$, with range $i \geq 0$.

To prove that $P(t)$ is the unique solution to the FKE, assume there exists another matrix $K(t)$ such that it satisfies FKE then

$$\frac{d}{dt}(K(t)e^{-tQ}) = M'(t)e^{-tQ} + M(t)\left(\frac{d}{dt}e^{-tQ}\right) = M(t)Qe^{-tQ} + M(t)(-Q)e^{-tQ} = 0$$

So $M(t)e^{-tQ}$ is constant $\forall t \geq 0$, hence $M(t) = P(t)$, $\forall t \geq 0$. Proofs of existence and uniqueness of BKE are analogous. □

If the state space S is finite the forward and backward equations have the same solution, and it is unique, hence the transition probabilities are defined by the matrix $P(t) = e^{tQ}$ that can be explicitly computed via diagonalization of e^{tQ}. We stick to the case where S is finite, as that is the general setting we will use in multi-state models.

To end with this section, we will present the definition of sojourn time in a CTMC.

Definition 1.2.13. *(sojourn time) For each non-absorbing state $i \in I$, we define the sojourn time as $S_i = \min\{t \geq 0 : X(t) \neq i, X(0) = i\}$, which is the time spent in i before jumping to another state. If a state $i \in I$ is absorbing (a state that once entered it cannot be left) then $S_i = \infty$.*

1.3 Mathematical analysis

Definition 1.3.1. *(Measurable space) Let $\mathcal{F}$ be a σ-algebra of a set X. Then the tuple $(X, \mathcal{F})$ is called a measurable space.*

Definition 1.3.2. *(Exponential matrix) Let $X \in \mathcal{M}(\mathbb{R}^{n\times n})$, its exponential denoted as e^X or $\exp(X)$ is defined as the $n \times n$ matrix given by the power series*

$$e^X = \sum_{k=0}^{\infty} \frac{X^k}{k!}$$

Proposition 1.3.1. *If X is a diagonal matrix with entries x_{ii} then $e^X = \{\exp(x_{ii})\}_{i=1}^n$. Otherwise, if X is a diagonalizable matrix then its exponential matrix can be explicitly computed as $e^X = Pe^D P^{-1}$ where D is a diagonal matrix and P a non-singular matrix.*

1.3.1 Product integral

This section follows, primarily, references [8] and [16]. Some results of probability theory can be elegantly expressed in the language of product integration. Product integrals will be used in MSM to construct likelihood functions.

Definition 1.3.3. *(Partition). A set $P = \{x_0, \dots, x_n\}$ is a partition of a closed set $A = [a, b] \subset \mathbb{R}$ if $x_0 = a < x_1 < \cdots < x_n = b$. The set of all possible partitions of the set A is denoted as $\mathcal{P}(A)$.*

Definition 1.3.4. *(Product integral). Let $g(x)$ be a Riemann integrable function over the interval $A = [a, b] \subset \mathbb{R}$, and a sequence of partitions $P_R \in \mathcal{P}(A)$ defined as $P_R = \{a = x_0, x_1, \dots, x_R = b\}$ with lengths $\Delta x_r = x_r - x_{r-1}$, $r = 1, \dots, R$. Assume that the sequence $\max_{0 \le r \le R} \Delta x_r \to 0$ when $R \to \infty$. The product integral of $g(x)$ over A is defined as*

$$\prod_{(a,b]} \{1 + g(x)dx\} := \lim_{R \to \infty} \prod_{r=1}^{R} \{1 + g(x_r)\Delta x_r\} \tag{1.5}$$

Proposition 1.3.2. *Product integral can be expressed via Riemann integral in the following manner*

$$\lim_{R \to \infty} \prod_{r=1}^{R} \{1 + g(x_r)\Delta x_r\} = \exp\left(\int_a^b g(x)dx \right) \tag{1.6}$$

Proof. I detail idea given by [8] in section 2.2.1, the authors state the equality of the proposition by just noting that a first order Taylor expansion of $\log(1 + x)$ jointly with $\exp(\log(1 + x))$ yields to the desired result, but beyond that, no details on the procedure are given. First, developing $\log(1 + g(x_r)\Delta x_r)$ as Maclaurin series

$$\log(1 + g(x_r)\Delta x_r) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \left(g(x_r)\Delta x_r\right)^k = \sum_{k=1}^{N} \frac{(-1)^{k+1}}{k} \left(g(x_r)\Delta x_r\right)^k + o(\Delta x_r^N)$$

An approximation of first order gives $\log(1 + g(x_r)\Delta x_r) = g(x_r)\Delta x_r + o(\Delta x_r)$, plugging it into the product integral expression:

$$\begin{aligned}
\lim_{R \to \infty} \prod_{r=1}^{R} \{1 + g(x_r)\Delta x_r\} &= \lim_{R \to \infty} \prod_{r=1}^{R} \exp\{\log(1 + g(x_r)\Delta x_r)\} = \lim_{R \to \infty} \prod_{r=1}^{R} \exp\{g(x_r)\Delta x_r + o(\Delta x_r)\} \\
&= \lim_{R \to \infty} \exp\left\{ \sum_{r=1}^{R} g(x_r)\Delta x_r + o(\Delta x_r) \right\} \overset{(a)}{=} \exp\left\{ \lim_{R \to \infty} \sum_{r=1}^{R} g(x_r)\Delta x_r + o(\Delta x_r) \right\} \\
&= \exp\left\{ \lim_{R \to \infty} \sum_{r=1}^{R} g(x_r)\Delta x_r + \lim_{R \to \infty} \sum_{r=1}^{R} o(\Delta x_r) \right\} \overset{(b),(c)}{=} \exp\left\{ \int_a^b g(x)dx \right\}
\end{aligned}$$

(a) Follows from continuity of the exponential function.

(b) As $g(x)$ is Riemann integrable over $[a, b]$ then the Riemann sum converges $\lim_{R \to \infty} \sum_{r=1}^{R} g(x_r)\Delta x_r = \int_a^b g(x)dx$.

(c) As $\Delta x_r \le \Delta x_{(R)}$ $\forall r = 1, \dots, R$, were $\Delta x_{(R)} := \max_{1 \le r \le R} \Delta x_r$, then by assumption $0 \le \Delta x_r \le \lim_{R \to \infty} \Delta x_{(R)} = 0$, and so $\Delta x_r \to_{R \to \infty} 0$.

For asymptotic equality to hold I must prove that $\lim_{R \to \infty} \sum_{r=1}^{R} \Delta x_r = 0$, to do so I will show $\left\{ \sum_{r=1}^{R} \Delta x_r \right\}_{R \ge 1}$ is a convergent sequence with limit 0.

$\lim_{\Delta x_r \to 0} \frac{o(\Delta x_r)}{\Delta x_r} = 0, \forall r = 1, \ldots, R \iff \forall \varepsilon > 0 \; \exists r_0 \in \mathbb{N} \text{ s.t } \forall r \geq r_0$

$$|o(\Delta x_r)| < \frac{\varepsilon}{b-a} |\Delta x_r|$$

We claim that $\forall R \geq r_0$:

$$\left| \sum_{r=1}^{R} o(\Delta x_r) \right| < \varepsilon$$

Indeed,

$$\left| \sum_{r=1}^{R} o(\Delta x_r) \right| \leq \sum_{r=1}^{R} |o(\Delta x_r)| < \sum_{r=1}^{R} \frac{\varepsilon}{b-a} |\Delta x_r| \overset{(a)}{=} \frac{\varepsilon}{b-a} \sum_{r=1}^{R} \Delta x_r \overset{(b)}{=} \frac{\varepsilon}{b-a}(b-a) = \varepsilon$$

(a) $\Delta x_r = x_r - x_{r-1} > 0$ because $\{x_r\}_{0 \leq r \leq R}$ is an increasing sequence of points.

(b) As $a = x_0 < \cdots < x_R = b$ is a partition and $\Delta x_r = x_r - x_{r-1}$ the sum $\sum_{r=1}^{R} \Delta x_r = (x_1 - x_0) + (x_2 - x_1) + \cdots + (x_{R-1} - x_{R-2}) + (x_R - x_{R-1}) = x_R - x_0 = b - a$ □

Chapter 2

Multi-state models

2.1 Motivation and some illustrative clinical studies

In a standard survival analysis study the experience of a patient may be modelled as a process with two states: dead and alive, and one possible transition from "alive" to "dead" state. Some studies might require to partition the "alive" state into more intermediate states. Multi-state models (MSM) are used to model the movement of patients among the various states. These models are suitable for longitudinal data, where subjects are observed over time and covariate information is collected in different occasions.

One could define multi-state models as a continuous time stochastic process allowing individuals to move among a finite number of states. States can be defined *ad hoc* based on study objectives such as: clinical symptoms (e.g bleeding episodes), biological markers (e.g CD4 T-lymphocyte cell count; serum immunoglobulin levels), scale of the disease (e.g stages of cancer or HIV infection), non-fatal complication in the course of the illness (e.g cancer recurrence). MSM have been used to explore the natural history of diseases and conditions as diverse as lung transplantation, cardiovascular diseases, politics, chronic myeloid leukemia, insurance, colon cancer, and psoriatic arthritis.

Some useful analysis multi-state models can carry out are

- *Competing risks.* MSM can be used to analyze a competing risks situation. Three states: alive, dead to illness, dead due to any of several causes, and two transitions: alive to dead, alive to dead due to any of several causes.

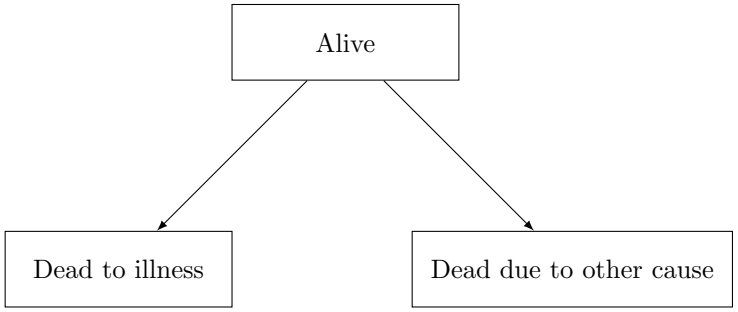


Figure 2.1: Competing risks as multi-state model

Source: Own elaboration

- *Increase statistical power.* In bioequivalence studies we are interested in comparing two or more treatments with respect to their ability to slow or prevent disease recurrence. An *illness-death*[1] model would be suitable for, even if the interest just lies on transition from state "disease-free after treatment" to "recurrence of disease". It will be necessary to consider state "death", otherwise we

[1] A MSM with three states: alive, recurrence of disease, death. Transitions from: alive to recurrence of disease, alive to death, recurrence of disease to death, are considered.

will tend to decrease power for testing treatment differences on the basis of reccurence. Treatment comparisons based on recurrence-free survival, instead than on death transitions, can have enhanced power if treatments have the same type of effect. Few theoretical studies on statistical power and sample size estimation for MSM have been carried out, though two simulation studies ([3],[17]) have shown that, under some assumptions, MSM can reduce sample size by half with respect to a standard survival analysis.

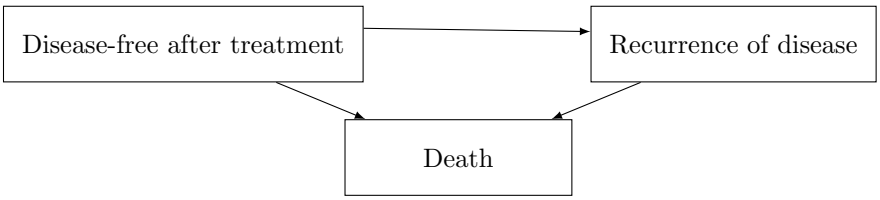


Figure 2.2: Illness-death model

Source: Own elaboration

- *Transition probabilities and sojourn times.* MSM gives two important extra tools: transition probabilities and sojourn times. Transition probabilities are the probabilities of moving from one state into another state, for instance the effectiveness of a treatment to prevent an illness might be tested through the transition probabilities, stratified by treatment, to a clinical relevant state. On the other hand, sojourn time is the amount of time an individual is expected to spend in a state before transitioning into another one. The effectiveness of a treatment to reduce the severity of the illness might be tested through sojourn times, instead of transition probabilities, as the clinical endpoint is to test the reduction of severity/time of the illness. Transition intensities provide the hazards for movement from one state to another, these functions are also used to determine the mean sojourn time in a given state and the number of individuals in different states at a certain time. MSM are useful when primary analysis is to estimate transition probabilities between states based on patient profile.

- *History of the disease.* Covariates may be incorporated in MSM through each transition intensity to explain differences between individuals. This tool helps to take into account that intermediate events can change the natural history of the disease progression, so that the role of some prognostic factors may not be the same after an individual reaches a certain state, prognostic factors associated with rates of death can be different in different transient (not absorbing) states.

- *Multiple endpoints.* MSM can take into account multiple survival endpoints at once, through transition intensities and sojourn times between and within states, respectively.

Throughout this chapter we will consider a generic individual on whom the following data has been previously collected:

- **Observed time of event**. Events or study endpoints are assumed to occur at a specific instant in time: infection, diagnosis of a disease, death, ... But only observed time of the endpoint is given, thus generating uncertainty about the exact time that the endpoint ocurred. We will assume the proper censoring type, of observed times, when needed. For each endpoint $k = 1, \ldots, R$ we will have observed times $t_1, \ldots, t_R$ for occurance of these.

- **Number of events ocurred**. For each endpoint we will have the number of times an individual has experienced it up to a given time. Counting process notation will be used to represent this information, for a fixed individual i and each endpoint k a counting process will be defined $\{N_{ik}(t), t \geq 0\}_{r=1}^{R}$, where $N_{ik}(t)$ denotes the count of ocurred endpoint k up to time t for individual i.

- **Factors**. Denote the status of an individual at a given time. No restriction on the number of factor's levels are imposed. Some examples: ranges for biomarkers (mutually exclusive), biological sex, treatment, dose. We will denote all this information for an individual i at observed time t with a vector $\mathbf{Z}_i(t)$.

- **Fixed-time covariates**. Covariates associated with individuals that are not likely to experience a change throughout the study, some examples: number of family members at home, year of birth, labour income. We will denote all this information for individual i with a vector $\mathbf{X}_i$.
- **Time-dependent covariates**. Covariates that might experience a change throughout the study, these can either be individual measures or external factors that might evolve over time. Some examples are: biomarkers, weight, air pollution, concomitant medication. We will denote all this information for an individual i at observed time t with a vector $\mathbf{X}_i(t)$.

In the next three subsections we will introduce the potential of MSM through some real studies that have carried out their analyses through these models, mostly on a clinical research setting. The aim of section 2.1.1 is to exemplify the use of hazard ratios and adjustment of covariates, whereas section 2.1.2 shows how transition probabilities are used to extract relevant information on cancer dynamics, finally section 2.1.3 brings an example on which states are defined on a creative way so as to study a biomarker's dynamics in HIV.

2.1.1 High risk and probability of progression to osteoporosis at 10 years in HIV-infected individuals: the role of PIs. [11]

HIV-infected people have a high risk of osteoporosis due to multiple factors, including the virus itself. In this retrospective longitudinal observational study, 875 patients were considered with at least 2 DXA scans[2] (3726 scans in total). Osteoporotic fractures still remain very infrequent and physicians rarely evaluate bone health, and when evaluated no certainty on HIV cause related is assessed. The objective of the study is to estimate the risk of progression to osteopenia or osteoporosis (worsening condition) among HIV-infected patients.

Patients are initially classified in three states based on their DXA scan ranges: normal bone mineral density, osteopenia, osteoporosis. A time-non-homogeneous[3] bidirectional multistate model based on these three states is fitted. All transitions are interval-censored. The corresponding MSM diagram for this study is given.

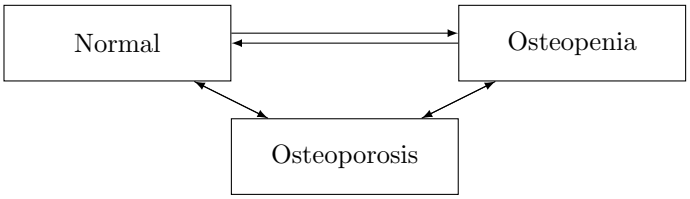


Figure 2.3: Diagram of a model used in progression to osteoporosis in HIV individuals

Source: Own elaboration

MSM allows for estimation of standard survival analysis inference stratifying by factors or covariates, but also transition probabilities and sojourn times. Hazards ratios stratified by sex and its 95% CIs are given for each possible transition between states.

[2] An imaging test that measures bone density (the amount of bone mineral contained in a certain volume of bone) by passing x-rays with two different energy levels through the bone.

[3] Transition probabilities not constant over time

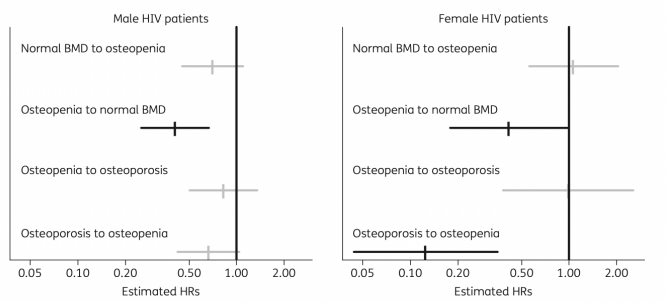


Figure 2.4: Estimated HRs and 95% CIs associated with age (>45 years versus $\leq$ 45 years) for model transitions. Lines in black indicate a greater likelihood of recovery from bone loss among younger HIV-infected patients. BMD, bone mineral density.

Source: Extracted from [11]

A hazard function of 1 implies an equal risk of transitioning from one state into another. A different pattern of risk is observed when stratifying by sex. Lower risk of developing osteopenia, when assessed as osteoporosis at screening, is observed for female HIV patients, as the 95% CI hazard ratio does not include 1. No other conclusive inferences can be drawn, only through this figure, as other CI include 1. Higher risk of developing osteopenia, when assessed as osteoporosis at screening, is observed for male HIV patients, as Higher risk of developing osteopenia, when assessed as normal bone mineral density at screening, is observed for female HIV patients.

The article also studies the risk of transition based on different Protease inhibitors given to patients.[4] are given to HIV/AIDS patients.

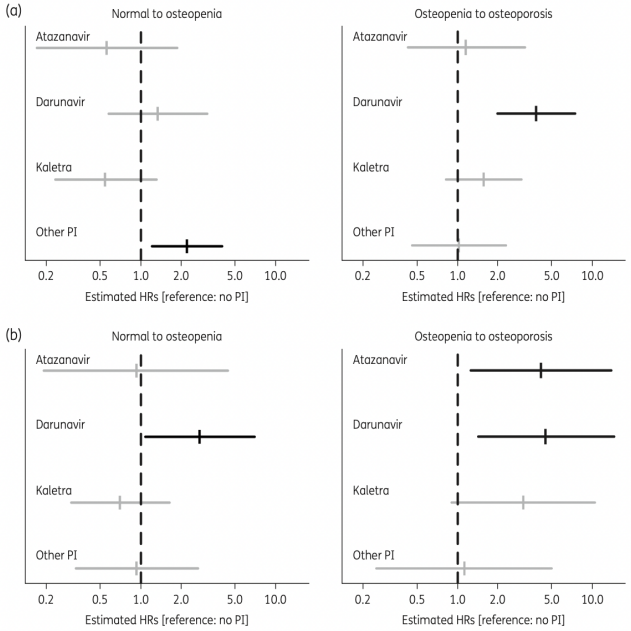


Figure 2.5: (a) Estimated HRs and 95% CIs associated with specific PIs among HIV-infected men. Left-hand panel: transition from normal bone mineral density to osteopenia. Right-hand panel: transition from osteopenia to osteoporosis. (b) Estimated HRs and 95% CIs associated with specific PIs among HIV-infected women. Left-hand panel: transition from normal bone mineral density to osteopenia. Right-hand panel: transition from osteopenia to osteoporosis.

Source: Extracted from [11].

From all HIV-infected females assessed at screening as normal bone mineral density, those taking

[4]Protease inhibitors (PIs) are medications that act by interfering with enzymes that cleave proteins. Some of the most well known are antiviral drugs widely used to treat HIV/AIDS and hepatitis C

Darunavir have a higher risk of experiencing osteopenia as those not taking any PI, as shown in hazard ratio CI (b). Likewise, from all HIV-infected males assessed at screening as normal bone mineral density, those taking Other PI have a higher risk of developing osteopenia. Thus, a higher risk of developing osteopenia is observed, for both male and female HIV patients, when taking Darunavir as PI compared to those not taking any PI.

Authors give two reasons of why a MSM approach has been proposed instead of other analysis:

- In clinical practice, patients are naturally classified by physicians according to their bone mineral density into: normal, osteopenia, osteoporosis. So the states arise naturally in this setting. As the article reports: "It is of medical interest to know, for example, what the probability is of a patient with normal bone mineral density of suffering from osteopenia within a certain period" [11].
- The course of bone mineral density is hardly captured using standard regression models, given that bone mineral density has its ups and downs: it might have improvement periods followed by deterioration periods.

2.1.2 Cohort Study on Performance Status Among Patients Diagnosed With Cancer. [18]

In observational studies on cancer patients, MSM can be used to study the progression of performance status over time. The cohort consisted on 11.342 patients diagnosed with cancer in Ontario, Canada, from 2007 to 2009. For this study 4 states are considered based on PPS score[5]: 1-stable (PPS score 70-100), 2-transitional state (PPS score 40-60), 3-end of life state (PPS score 10-30), 4-deceased. Patients are assessed by a physician only at periodic clinic or home visits, in these cases, the exact times of state-to-state transitions, other than death, are interval-censored (i.e transition is only known to have occurred within a bounded time interval, usually between assessments). A bidirectional time-homogeneous Markov MSM is fitted. The following MSM diagram is considered.

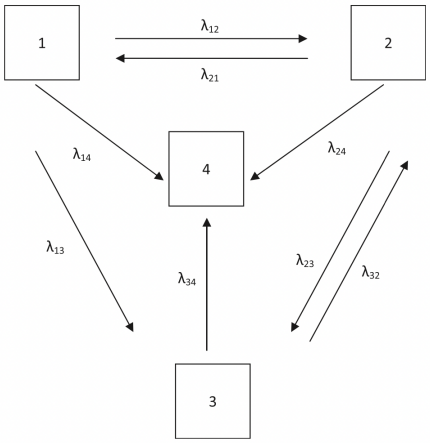


Figure 2.6: Underlying 4-state model for examining disease progression among cancer patients using the Palliative Performance Scale (PPS), PPS Cohort Study, Ontario, Canada, 2007–2009. (1—stable state; 2—transitional state; 3—end-of-life state; 4—deceased).

Source: Extracted from [18].

Each λ_{ij} represent the transition intesity function between state $i \to j$. The main study goal is to estimate these transition intensity functions so as to provide insight into the nature of the progression of performance status of cancer patients over time.

[5]Palliative Performance Scale (PPS) is a validated and reliable tool used to assess a patient's functional performance and to determine progression toward end of life.

Month and State[a]	Maximum Likelihood Estimate			
	State 1	State 2	State 3	State 4
1 month				
State 1	0.945	0.044	0.003	0.008
State 2	0.109	0.596	0.053	0.242
State 3	0.005	0.052	0.207	0.735
State 4	0.0	0.0	0.0	1.0
6 months				
State 1	0.750	0.087	0.009	0.154
State 2	0.216	0.068	0.008	0.708
State 3	0.019	0.008	0.001	0.972
State 4	0.0	0.0	0.0	1.0

Table 2.1: Estimated 1-Month and 6-Month Transition Probabilities Among Cancer Patients in the PPS Cohort Study, Ontario, Canada, 2007–2009.

Source: Extracted from [18].

Figure 2.1 gives the transition probability matrix of the MSM stratified by month, for absorbing states the probability of transition into another state is 0, as once the individual enters 4-deceased state it remains on it forever. No matter which month we look at, once a patient is diagnosed in 3-end-of-life stage, the probabilities of improving condition is quite low as the chances of a patient to transient from end-of-life stage to either stable state or transitional state, is really low (less than 1%). A patient in the transitional state (2) has chance of experiencing an improvement in performance status at the end of 6 months, since at month 1, 60% of patients that were on the transitional state remain on it after first month, once they reach the sixth month a pattern is observed: 70% of patients that were in transitional state die, 22% improve their condition to the stable state.

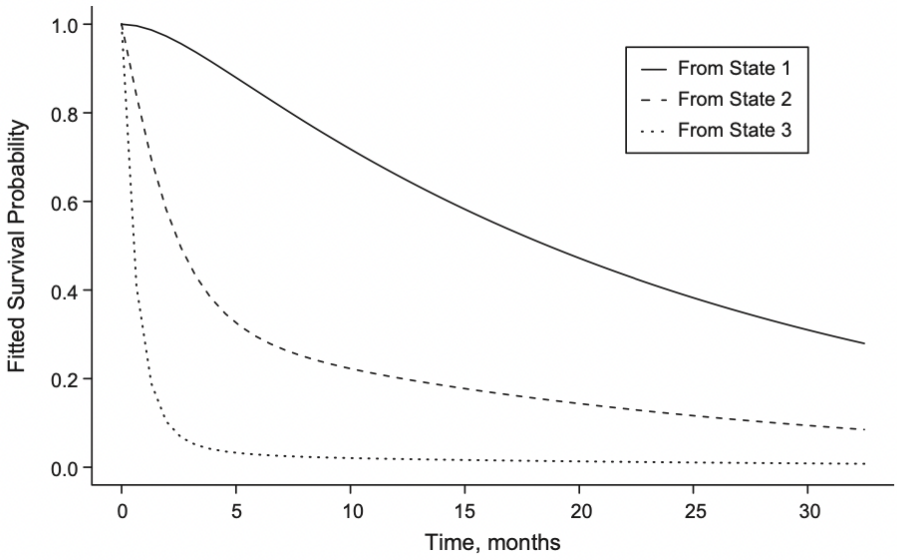


Figure 2.7: Estimated survival probability over time (months) from each nonabsorbing state to death among cancer patients in the PPS Cohort Study, Ontario, Canada, 2007–2009. (PPS, Palliative Performance Scale).

Source: Extracted from [18].

Figure 2.7 shows that the survival probability of an end-of-life stage patient drops drastically after first month, so waiting until a patient reaches the end-of-life state may be too late for discussing hospice or palliave-care, since patients do not remain on this state for more than a month.

2.1.3 Viral Load Dynamics in Individuals with HIV Infection [8]

The Canadian Observational Cohort on HIV (CANOC) is composed of several Canadian cohorts of HIV-positive persons who initiated combination antiretroviral therapy. Selection criteria is based on achieving initial viral suppression, such individuals are followed up at visits with time windows of 3/4 months. In each visit some HIV-related biomarkers are measured such as: CD4, CD8, viral load count, blood lipid level. Furthermore, for each patient covariates related to: AIDS-defining illnesses, cardiovascular events, diagnoses of cancer, death, are collected. This data, jointly with a MSM approach, gives insight into dynamics of HIV disease infection and its associated factors, the effectiveness of treatment and optimal strategies for patient management. The primary clinical endpoint is the study of the dynamics of HIV biomarkers, states will be defined as ranges of those biomarkers (based on clinical significance). The corresponding MSM's diagram of this study is shown below.

Figure 2.8: Diagram of a model used in HIV viral load study.

Source: Own elaboration.

2.2 Counting processes and multistate models

Event history analysis deals with ocurrence of events over time. A generic individual will be considered, for the sake of simplicity we will avoid indexing for individuals. As mentioned in section 2.1 the process $\{N_k(t), t \geq 0\}$ is a counting process, a vector of all these processes gives a multivariate counting process $\{N(t), t \geq 0\}$ where $N(t) = (N_1(t), \ldots, N_R(t))'$.

Denote $\mathcal{H}_N(t) = \{N(s), 0 \leq s \leq t\}$ and $\mathcal{H}_N(0^-) = \emptyset$ the filtration of the counting process $N(t)$ over $[0, t]$, and $k = 1, 2, \ldots, R$ the events.

Definition 2.2.1. *(Intensity function for event k). The intensity function for events of type k is defined as*

$$\lambda_k(t|\mathcal{H}_N(t^-)) = \lim_{\Delta t \to 0^+} \frac{\mathbb{P}(\Delta N_k(t) = 1|\mathcal{H}_N(t^-))}{\Delta t} \tag{2.1}$$

for $t \geq 0$ where $\Delta N_k(t) = N_k(t + \Delta t^-) - N_k(t^-)$.

Two or more events cannot occur simultaneously, since we only consider mutually exclusive states, then the intensity functions $k = 1, \ldots, R$ fully specify the multivariate event process. When considering MSM with a state space $S = \{1, \ldots, K\} \subset \mathbb{N}$ intensity functions between states are defined similarly. Another useful counting process is $\{Z(t), t \geq 0\}$ which counts the state a subject is at a given time t.

Definition 2.2.2. *(Intensity function from state $k \to l$). The intensity function for transition from state $k \to l$ is defined as*

$$\lambda_{kl}(t|\mathcal{H}_Z(t^-)) = \lim_{\Delta t \to 0^+} \frac{\mathbb{P}(Z(t + \Delta t^-) = l|Z(t^-) = k, \mathcal{H}_Z(t^-))}{\Delta t} \tag{2.2}$$

for $k \neq l$ and $Z(t)$ denotes the state occupied at time $t \geq 0$.

The filtration $\mathcal{H}(t)$[6] is of key importance, it defines the whole MSM: its stochastic dependence with past information and functional relationship with covariates. We can expand the meaning of the filtration so that it includes all relevant covariates; considering a fixed covariate vector $\mathbf{X}$ then $\mathcal{H}_Z(t) = \{Z(s) : 0 \leq s \leq t, \mathbf{X}\}$, from this we can make assumptions on how the process behaves with respect to the filtration $\mathcal{H}_Z(t)$

- **Stochastic dependence**, which means how past information about the process interferes with present information of it, i.e how previous states can interfere with current and future ones. It all sums

[6] As a clarification, we refer to $\mathcal{H}(t)$ as either $\mathcal{H}_N(t)$ or $\mathcal{H}_Z(t)$ interchangeably.

up on how the process behaves with respect to the filtration $\mathcal{H}(t)$, different assumptions can be made out of this. Due to a large amount of literature and straightforward theoretical results, most notably studied processes in MSM framework are Markov processes and semi-Markov processes.

  - *Markov processes.* Suitable when the dependence on the history is only through current state occupied, and so previous states occupied by the individual do not influence transition intensities, and hence do not influence anything in MSM (transition probabilities, sojourn times, inference, etc.). Formally, for a MSM to be a Markov process it suffices that $\lambda_{kl}(t|\mathcal{H}_Z(t^-)) = \lambda_{kl}(t|\mathbf{X})$. Modulated Markov processes are Markov processes that incorporate time dependent covariates $\mathbf{X}(t)$ so that: $\lambda_{kl}(t|\mathcal{H}_Z(t^-)) = \lambda_{kl}(t|\mathbf{X}(t))$.
  - *Semi-markov processes.* Current state just depends on the time since entry to itself. Formally, for a MSM to be a semi-Markov process $\lambda_{kl}(t|\mathcal{H}_Z(t^-)) = \lambda_{kl}(B(t)|\mathbf{X})$ where $B(t)$ is the time since first entry to the current state k. Modulated semi-markov processes are semi-markov processes that incorporate time dependent covariates so that: $\lambda_{kl}(t|\mathcal{H}_Z(t^-)) = \lambda_{kl}(B(t)|\mathbf{X}(t))$. Note that a semi-Markov process is non time-homogenous by definition.
  - *non-Markov processes*, which is any process that does not exhibit the Markov property. For instance, in a clinical study, a patient's condition can be permanently modified once it occupies a certain state, and thus a Markov assumption would be too strong to assume.

- **Functional dependence with respect to covariates**. Introduction of covariates in these models allows for prediction of probabilities based on individual's profile, a regression-type functional dependence between covariates and transition intensity functions is assumed. As usual, $\mathbf{X}$ is a $(p\times 1)$-dimensional vector of fixed covariates coming from a sample space $\mathcal{X} \subseteq \mathbb{R}^p$ and β_{kl} is a $(p\times 1)$-dimensional vector of coefficients of the regression from state $k \to l$ with parameter space $\Theta_{kl} \subseteq \mathbb{R}^p$, and a non-negative function $g : \mathcal{X} \times \Theta_{kl} \to \mathbb{R}^+$, covariates can be introduced in two different ways:

  - *Additive form*, $\lambda_{kl}(t|\mathcal{H}(t^-), \mathbf{X}) = \lambda_{kl}(t|\mathcal{H}(t^-)) + g(\mathbf{X}; \beta_{kl})$, which might have some inconveniences like to restrict the parameter space so as to get a non-negative transition intensity function.
  - *Multiplicative form*, $\lambda_{kl}(t|\mathcal{H}(t^-), \mathbf{X}) = \lambda_{kl}(t|\mathcal{H}(t^-))g(\mathbf{X}; \beta_{kl})$, a common approach is to set $g(\mathbf{X}; \beta_{kl}) = \exp(\beta'_{kl}\mathbf{X})$ that jointly with a Markov assumption yields to a proportional hazards model. With the same approach for g and assuming a process that behaves as a time-homogeneous Markov process, a Cox proportional hazards model can be fitted.

Example 1) Functional form of a regression from state $k \to l$ of time-homogeneous Markov process with a multiplicative functional dependence with respect to covariates $g(\mathbf{X}; \beta_{kl}) = \exp(\beta'_{kl}\mathbf{X})$:

$$\lambda_{kl}(t|\mathcal{H}_Z(t^-)) = \lambda_{kl|\mathbf{X}} = \lambda_{kl,0} \exp(\beta'_{kl}\mathbf{X}) = \exp(\log(\lambda_{kl,0}) + \beta'_{kl}\mathbf{X})$$

Where $\lambda_{kl,0}$ is known as *baseline intensity*, which is the intensity for a subject with profile $\mathbf{X} = 0$

Example 2) Functional form of a regression from state $k \to l$ of a modulated semi-Markov process with additive functional depence with respect to covariates $g(\mathbf{X}(t); \beta_{kl}) = \beta'_{kl}\mathbf{X}(t)$

$$\lambda_{kl}(B(t)|\mathcal{H}_Z(t^-)) = \lambda_{kl}(B(t)|\mathbf{X}(t)) = \lambda_{kl,0}(B(t)) + \beta'_{kl}\mathbf{X}(t)$$

Where $\lambda_{kl,0}(B(t))$ is the *baseline intensity* for an individual with profile $\mathbf{X}(t) = 0$ for all t.

MSM can get mathematically very challenging, even intractable, once one drops the assumption of Markov process, for this reason, from now on we will devote the theory and applications of MSM driven by an homogeneous Markov process. Likewise, a multiplicative functional form will be assumed, hence a Cox proportional hazards model will be fitted for transition intensities.

2.3 Intensity Functions and Counting Processes

In this section counting process notation for multi-state models is presented, as introduced in section 2.2 two counting processes can be defined for a MSM: the state occupied at time t represented by $\{Z(t), t \geq 0\}$, and the number of transitions occurred at time t represented by $\{N(t), t \geq 0\}$. To fix ideas, so far we have defined three stochastic processes for a MSM:

- A continuous time stochastic process for the transition intensities as either: Markov process, semi-Markov, non-Markov.
- A counting process for the state occupied at time t represented by $\{Z(t), t \geq 0\}$.
- A counting process for the number of transitions occurred at time t represented by $\{N(t), t \geq 0\}$.

It turns out that there is no need to use two counting processes to define MSM, it suffices to take only one of these two counting processes. In fact, we will see that the two counting processes defined are equivalent, which means that the process $N(t)$ gives the same information as the process $Z(t)$ at every timepoint t, and all MSM can be reduced to two stochastic processes: a continuous time stochastic process and a counting process. There is no need to define two intensity functions for each of the counting processes, as both processes are equivalent. The aim of this section will be to prove Proposition 2.3.1 which gives the desired result.

We define an extra counting process $N_{kl}(t)$ that counts the number of instantaneous transitions from state $k \to l$ over a time period $[0, t]$.

Two key quantities are the core of why $N_k(t)$ is equivalent to $Z(t)$:

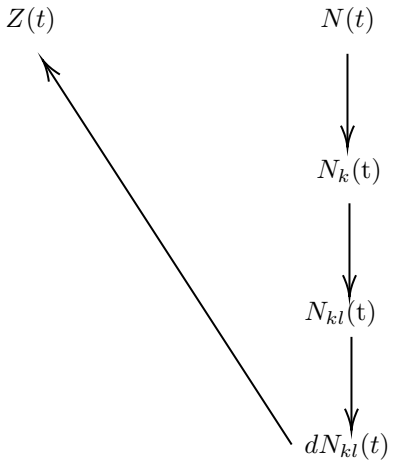
$$\Delta N_{kl}(t) = N_{kl}(t + \Delta t^-) - N_{kl}(t^-) \tag{2.3}$$

$$dN_{kl}(t) = \lim_{\Delta t \to 0^+} \Delta N_{kl}(t) \tag{2.4}$$

Equality (2.3)[7] gives the number of transitions $k \to l$ over a time interval $[t, t + \Delta t)$, and (2.4) is related to the number of transitions from $k \to l$ at time t. Since $N_{kl}(t)$ is a counting process, it is right continuous ((2.4) exists) and has jumps of size 1, for some fixed t then, $dN_{kl}(t)$ is a Bernoulli random variable that can either take a value of 1 or 0. It is clear that if $dN_{kl}(t) = 1$ then a transition from $k \to l$ at time t ocurred and so $Z(t+\Delta t^-) = l$ given that $Z(t^-) = k$. The vector $dN_k(t) = (dN_{kl}(t), l \in S \setminus \{k\})'$ contains the information on all transitions at time t from state k to any other state in the state space $S \setminus \{k\}$. If $\sum_{l \in S \setminus \{k\}} dN_{kl}(t) = 0$ no transition occurred from state $k \to l$ at time t, so everything that was in state k at t^- remains on it at time t. Likewise, if $\sum_{l \in S \setminus \{k\}} dN_{kl}(t) = 1$[8] a transition from k to another state $S \setminus \{k\}$ occurred, and the element $dN_{kl}(t) = 1$ denotes the state to which the transition occurred.

Coming back to the initial processes $N_k(t)$ and $Z(t)$, we have already seen that there's an intrinsic relationship between processes $Z(t)$ and $dN_{kl}(t)$, but there's still no link on how $N_k(t)$ and $Z(t)$ might be equivalent. In fact, $N_k(t)$ can be expressed via $N_{kl}(t)$, $N_k(t) = (N_{kl}(t), l \in S \setminus \{k\})'$, which is a clearer definition given that $N_{kl}(t) = \int_0^t dN_{kl}(s)$ is the cumulative number of transitions from state k to all other possible states $S \setminus \{k\}$ over $[0, t]$.

To fix ideas, the next diagram aims to clarify to the reader the motivation of the definitions introduced:



[7] t^- denotes the time right before t.

[8] At an exact time of t if a transition occurred this sum must be equal to one, because only one transition can occur over an infitesimal time interval.

We have decomposed the process $N(t)$ into smaller subprocesses to show that there is a potential link between $N(t)$ and $Z(t)$. This link will be given by the process $dN_{kl}(t)$, since it contains the same information as $Z(t)$ (given the value of $dN_{kl}(t)$ at every timepoint t we can infer the value of $Z(t)$ at every timepoint t).

An alternative way of representing $Z(t)$ is by $N(t) = (N_1(t), \dots, N_R(t))'$, which records the number of all transitions occurring over $[0, t]$, with the filtration $\{N(s), 0 \leq s \leq t; Z(0)\}$ which incorporates the information on the first state occupied, note that this information needs to be added up since is not naturally included in the natural filtration of $N(s)$ but indeed is in the $Z(t)$ one. Formally, all these previous reasoning summarizes on the following proposition.

Proposition 2.3.1. *Intensity function for event $k \to l$ can be expressed via transition intensity functions between states as*

$$\lim_{\Delta t \to 0^+} \frac{\mathbb{P}(\Delta N_{kl}(t) = 1 | \mathcal{H}(t^-))}{\Delta t} = Y_k(t^-)\lambda_{kl}(t|\mathcal{H}(t^-)) \tag{2.5}$$

where $Y_k(t) = I_{\{Z(t)=k\}}(t)$ indicator that the process is in state k at t

Proof. If $\Delta N_{kl}(t) = 1$ then a transition occurred over $[t, t+\Delta t)$ which, by definition, implies $Z(t+\Delta t^-) = l$, the reverse is also true. For each $t > 0$, define the events $A_t = \{\Delta N_{kl}(t) = 1\}$ and $B_t = \{Z(t+\Delta t^-) = l\}$ then:

$$\mathbb{P}(A_t|B_t) = \mathbb{P}(B_t|A_t) = 1, \forall t > 0$$

So events A_t and B_t are equivalent under $\mathbb{P}$. Expressing $\lambda_{kl}(t|\mathcal{H}(t^-))$ in terms of A_t:

$$\begin{aligned}\lambda_{kl}(t|\mathcal{H}(t^-)) = \lim_{\Delta t \to 0^+} \frac{\mathbb{P}(Z(t+\Delta t^-) = l | Z(t^-) = k, \mathcal{H}(t^-))}{\Delta t} &= \lim_{\Delta t \to 0^+} \frac{Y_k(t^-)\mathbb{P}(B_t|\mathcal{H}(t^-))}{\Delta t} \\ &= Y_k(t^-) \lim_{\Delta t \to 0^+} \frac{\mathbb{P}(A_t|\mathcal{H}(t^-))}{\Delta t}\end{aligned}$$

Second equality follows from:

$$\mathbb{P}(B_t|Z(t^-) = k, \mathcal{H}(t^-)) = \begin{cases} 0 & , Z(t^-) \neq k \\ \mathbb{P}(B_t|\mathcal{H}(t^-)) & , Z(t^-) = k \end{cases}$$

Which can be rewritten using an indicator function $Y_k(t^-) = I_{\{Z(t^-)=k\}}(t^-)$. To prove the desired equality we need to swap the indicator function $Y_k(t^-)$, it suffices to prove that if $Y_k(t^-) = 0$ then $\lim_{\Delta t \to 0^+} \frac{\mathbb{P}(A_t|\mathcal{H}(t^-))}{\Delta t} = 0$. If $Y_k(t^-) = 0$ then previous state occupied is not k:

$$A_t = \{\Delta N_{kl}(t) = 1\} = \{N_{kl}(t+\Delta t^-) - N_{kl}(t^-) = 1\} \overset{(a)}{=} \{N_{kl}(t+\Delta t^-) = 1\} \overset{(b)}{=} \emptyset$$

- (a), if $Y_k(t^-) = 0$ then $N_{kl}(t^-) = 0$, as at time t^- we were not in state k and so no transition could occur from $k \to l$, by definition $N_{kl}(t^-) = 0$ must hold.
- (b), if $Y_k(t^-) = 0$ then $\mathbb{P}(Z(t+\Delta t^-) = l | Z(t^-) = k, \mathcal{H}(t^-)) = 0$, which implies no transition can happen over time interval $[t, t+\Delta t)$ and so, by definition, $N_{kl}(t+\Delta t^-) = 0$.

And so when $Y_k(t^-) = 0$ we must have $A_t = \emptyset$, $\forall t > 0$. Then

$$\lim_{\Delta t \to 0^+} \frac{\mathbb{P}(A_t|\mathcal{H}(t^-))}{\Delta t} = 0$$

Which means we can swap the indicator to the other side of the equation as desired.

$$\lim_{\Delta t \to 0^+} \frac{\mathbb{P}(A_t|\mathcal{H}(t^-))}{\Delta t} = Y_k(t^-)\lambda_{kl}(t|\mathcal{H}(t^-))$$

□

Proposition 2.3.1 implies that only one counting process, either $N_{kl}(t)$ or $Z(t)$, suffices to define a MSM, since $N(t)$ can be expressed via $N_{kl}(t)$, $N(t)$ is also equivalent to $Z(t)$.

2.4 Mean, total, and sojourn time distributions

We treat a general stochastic process for the transition intensity functions, so time-homogeneous Markov Multi-state models will be a particular case from the general theory developed in this section. Some of the questions and answers the reader will face in this section are:

- What is a sojourn time in a MSM?
- How much time does it take to jump from one state to another?
- How can we validate results of a MSM?
- What are the total and mean sojourn times? What are they used for?

We let $W_k^{(r)}$ denote the sojourn time in state $k \in S$, on the r-th occasion it is entered, intuitevely $W_k^{(r)}$ is the amount of time a subject spends on state k at the r-th entrance on that state. Clearly, sojourn times are infinite for absorbing states, so the following results are just true for non-absorbing states.

Theorem 2.4.1. *(**Problem 2.11** from [8]) Let $T_k^{(r)}$ and $V_k^{(r)}$ be the entry and exit times for the r-th visit to state k for a multistate process with intensity functions $\lambda_{kl}(t|\mathcal{H}(t^-))$, define $W_k^{(r)} = V_k^{(r)} - T_k^{(r)}$ as the r-th sojourn time in state k, then $W_k^{(r)} \sim \exp(1)$.*

Proof. I will first prove, by product integration, that[9]

$$\mathbb{P}\left(W_k^{(r)} > w \mid t_k^{(r)}, \mathcal{H}(t_k^{(r)-})\right) = \exp\left\{-\int_{t_k^{(r)}}^{t_k^{(r)}+w} \sum_{l \in S\setminus\{k\}} \lambda_{kl}(u \mid \mathcal{H}(u^-))du\right\}$$

The main idea is to use conditional probability, on a sequence of partitions, to express the probability as a product and then using product integration to get the desired result.

$$\begin{aligned}\mathbb{P}\left(W_k^{(r)} > w \mid t_k^{(r)}, \mathcal{H}(t_k^{(r)-})\right) &= \mathbb{P}\left(V_k^{(r)} - T_k^{(r)} > w \mid t_k^{(r)}, \mathcal{H}(t_k^{(r)-})\right) = \mathbb{P}\left(V_k^{(r)} > T_k^{(r)} + w \mid t_k^{(r)}, \mathcal{H}(t_k^{(r)-})\right) \\ &= \mathbb{P}\left(V_k^{(r)} > t_k^{(r)} + w \mid \mathcal{H}(t_k^{(r)-})\right)\end{aligned}$$

Now define a partition for each r visit $P_M^{(r)} \in \mathcal{P}\left(\left[t_k^{(r)}, t_k^{(r)} + w\right]\right)$ with $u_{(M)}^{(r)} := \max_{1\le m\le M} \Delta u_m^{(r)} \to 0$ as $M \to \infty$, such that $t_k^{(r)} = u_0^{(r)} < u_1^{(r)} < \cdots < u_M^{(r)} = t_k^{(r)} + w$

$$\begin{aligned}\mathbb{P}\left(V_k^{(r)} > t_k^{(r)} + w \mid \mathcal{H}(t_k^{(r)-})\right) &\overset{(a)}{=} \lim_{M\to\infty} \prod_{m=1}^{M} \mathbb{P}\left(V_k^{(r)} > u_m^{(r)} \mid V_k^{(r)} > u_{m-1}^{(r)}, \mathcal{H}(u_{m-1}^{(r)-})\right) \\ &= \lim_{M\to\infty} \prod_{m=1}^{M} \left(1 - \mathbb{P}\left(V_k^{(r)} < u_m^{(r)} \mid V_k^{(r)} > u_{m-1}^{(r)}, \mathcal{H}(u_{m-1}^{(r)-})\right)\right)\end{aligned}$$

(a) Conditional probability, note that $\mathbb{P}(V_k^{(r)} > u_{(0)}^{(r)}) = 1$, $\forall k \in S$, $\forall r = 1, \ldots, R$.

The idea is now to understand what $\mathbb{P}(V_k^{(r)} < u_m^{(r)} \mid V_k^{(r)} > u_{m-1}^{(r)}, \mathcal{H}(u_{m-1}^{(r)}))$ truly means, this is probability that the exit time is less than $u_m^{(r)}$ knowing that we were still in state k when time was greater than $u_{m-1}^{(r)}$, we can go back to a counting process notation to express this probability. Exit time will be less than $u_m^{(r)}$ if $Z(u_m^{(r)}) = l$, for any $l \in S\setminus\{k\}$ when $Z(u_{m-1}^{(r)}) = k$, then

$$\mathbb{P}\left(V_k^{(r)} < u_m^{(r)} \mid V_k^{(r)} > u_{m-1}^{(r)}, \mathcal{H}(u_{m-1}^{(r)-})\right) = \sum_{l \in S\setminus\{k\}} \mathbb{P}\left(Z(u_m^{(r)}) = l \mid Z(u_{m-1}^{(r)}) = k, \mathcal{H}(u_{m-1}^{(r)-})\right)$$

[9] For the sake of simplicity, an abuse of notation will be used throughout this section. We refer to $\mathbb{P}\left(W_k^{(r)} > w \mid t_k^{(r)}, \mathcal{H}(t_k^{(r)-})\right)$ as $\mathbb{P}\left(W_k^{(r)} > w \mid T_k^{(r)} = t_k^{(r)}, \mathcal{H}(t_k^{(r)-})\right)$

Rewriting the probabilities we get

$$\lim_{M\to\infty}\prod_{m=1}^{M}\left(1-\sum_{l\in S\setminus\{k\}}\mathbb{P}\left(Z(u_m^{(r)})=l\mid Z(u_{m-1}^{(r)})=k,\mathcal{H}(u_{m-1}^{(r)-})\right)\right)$$
$$\overset{(b)}{=}\lim_{M\to\infty}\prod_{m=1}^{M}\left(1-\sum_{l\in S\setminus\{k\}}\lambda_{kl}(u_m^{(r)}\mid\mathcal{H}(u_{m-1}^{(r)}))\Delta u_m^{(r)}\right)$$

(b) By definition 2.2, in the discrete case, $\lambda_{kl}(u_m^{(r)}|\mathcal{H}(u_{m-1}^{(r)})) = \frac{\mathbb{P}(Z(u_m^{(r)})=l|Z(u_{m-1}^{(r)})=k,\mathcal{H}(u_{m-1}^{(r)}))}{\Delta u_m^{(r)}}$.

By using Proposition 1.3.2 the result follows:

$$\lim_{M\to\infty}\prod_{m=1}^{M}\left(1-\sum_{l\in S\setminus\{k\}}\lambda_{kl}(u_m^{(r)}\mid\mathcal{H}(u_{m-1}^{(r)}))\Delta u_m^{(r)}\right)=\exp\left\{-\int_{t_k^{(r)}}^{t_k^{(r)}+w}\sum_{l\in S\setminus\{k\}}\lambda_{kl}(u\mid\mathcal{H}(u^-))du\right\}$$

Setting $v := \varphi(w; t_k^{(r)})$, since $t_k^{(r)}$ is known and w is non-random then v is non-random, were $\varphi(\cdot,\cdot)$ is a primitive of the above integral, it's clearer than $W_k^{(r)} \sim \exp(1)$:

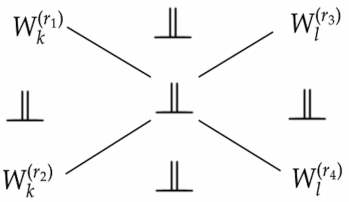
$$f_{W_k^{(r)}}(v) = \exp\{-v\}, v > 0$$

□

Intuitevely, last theorem tells us that the time it takes to jump from one state to another is exponentially distributed with mean 1. The following theorem states that the time we spend on a state until we jump to another one is independent from how many times we have entered that state, and the time a subject has spent on other states.

Theorem 2.4.2. *(**Problem 2.11** from [8]) Distinct $W_k^{(r)}$ $r = 1, 2, \ldots$, $k \in S$ are mutually independent.*

Proof. The aim goal will be to prove the following diagram:

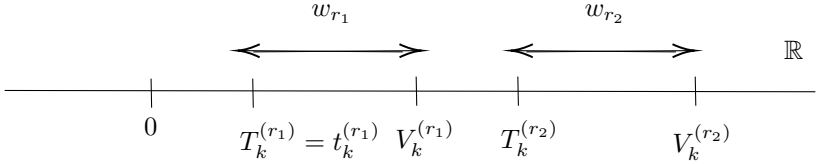


Where $\perp\!\!\!\perp$ denotes the independence between the random variables, $\{r_i\}_{1\le i\le 4}$ some arbitrary visits with $r_1 \neq r_2$ and $r_3 \neq r_4$, and $l, k \in S$ with $l \neq k$. To simplify notation define the set $\xi_k^r = \{W_k^{(r)} \geq w_r\}$, where $W_k^{(r)} = V_k^{(r)} - T_k^{(r)}$. I start by proving that $\forall k \in S$ and $\forall r_1, r_2 \in \mathbb{N}$ with $r_1 \neq r_2$: $W_k^{(r_1)} \perp\!\!\!\perp W_k^{(r_2)}$, by showing:

$$\frac{\mathbb{P}(\xi_k^{r_1}\cap\xi_l^{r_2}|t_k^{(r_1)},\mathcal{H}(t_k^{(r_1)-}))}{\mathbb{P}(\xi_l^{r_2}|t_k^{(r_1)},\mathcal{H}(t_k^{(r_1)-}))}=\mathbb{P}(\xi_k^{r_1}|t_k^{(r_1)},\mathcal{H}(t_k^{(r_1)-})) \tag{2.6}$$

By cases:

- Assume $r_1 < r_2$. Then the sets: $\xi_k^{r_1}$, $\xi_k^{r_2}$, can be thought as:



Note that $T_k^{(r_2)}$ is independent with respect to the filtration $\mathcal{H}(t_k^{(r_1)-})$ because $r_1 < r_2$, and $V_k^{(r_2)} > T_k^{(r_2)} \geq V_k^{(r_1)} > T_k^{(r_1)}$ must hold under the assumption.

$$\mathbb{P}(\xi_k^{r_1} \cap \xi_k^{r_2} | t_k^{(r_1)}, \mathcal{H}(t_k^{(r_1)-})) = \mathbb{P}(V_k^{(r_1)} - T_k^{(r_1)} \geq w_{r_1}, V_k^{(r_2)} - T_k^{(r_2)} \geq w_{r_2} | t_k^{(r_1)}, \mathcal{H}(t_k^{(r_1)-}))$$
$$\overset{(a)}{=} \mathbb{P}(V_k^{(r_2)} \geq t_k^{(r_1)} + w_{r_1} + w_{r_2} | \mathcal{H}(t_k^{(r_1)-})) \overset{(b)}{=} \exp\{-(w_{r_1} + w_{r_2})\}.$$

(a) By the following chain of inequalities: $V_k^{(r_2)} - T_k^{(r_2)} \geq w_{r_2} \iff V_k^{(r_2)} \geq T_k^{(r_2)} + w_{r_2}$ using that $T_k^{(r_2)} \geq V_k^{(r_1)}$ and imposing the set $\xi_k^{r_1} = \{W_k^{(r_1)} \geq w_{r_1}\} = \{V_k^{(r_1)} \geq t_k^{(r_1)} + w_{r_1}\}$ we get the intersection $\xi_k^{r_1} \cap \xi_k^{r_2}$ and the desired inequality.
(b) $V_k^{(r_2)} - t_k^{(r_1)}$ is a sojourn time, by Theorem 2.4.1: $V_k^{(r_2)} - t_k^{(r_1)} \sim \exp(1)$.
Plugging it into (2.6):

$$\frac{\mathbb{P}(\xi_k^{r_1} \cap \xi_k^{r_2} | t_k^{(r_1)}, \mathcal{H}(t_k^{(r_1)-}))}{\mathbb{P}(\xi_k^{r_2} | t_k^{(r_1)}, \mathcal{H}(t_k^{(r_1)-}))} = \frac{\exp\{-(w_{r_1} + w_{r_2})\}}{\exp\{-w_{r_2}\}} = \exp\{-w_{r_1}\} = \mathbb{P}(\xi_k^{r_1} | t_k^{(r_1)}, \mathcal{H}(t_k^{(r_1)-}))$$

- Assume $r_1 > r_2$. The event $\xi_k^{r_2}$ is known with respect to $\mathcal{H}(t_k^{(r_1)-})$, since $\xi_k^{r_1} = \{V_k^{(r_1)} - T_k^{(r_1)} \geq w_{r_1}\}$ does not depend on $V_k^{(r_2)}$ nor $T_k^{(r_2)}$ then necessarily:

$$\mathbb{P}(\xi_k^{r_1} | \xi_k^{r_2}, t_k^{(r_1)}, \mathcal{H}(t_k^{(r_1)-})) = \mathbb{P}(\xi_k^{r_1} | t_k^{(r_1)}, \mathcal{H}(t_k^{(r_1)-}))$$

The next step is to prove that $\forall k \in S$, $\forall l \in S \backslash \{k\}$, $\forall r_1, r_2 \in \mathbb{N}$ (can be equal): $W_k^{(r_1)} \perp\!\!\!\perp W_l^{(r_2)}$. Again we split the proof by cases based on which relationship holds between states k and l, remember that k and l are non-absorbing states. From now on we define the sets $\xi_s^r = \{W_s^{(r)} \geq w_{r,s}\}$, $\forall s \in S$ and $\forall r \in \mathbb{N}$. For every non-absorbing state $k \in S$ define the sets:

$$D_{k \leftrightarrow l} = \{l \in S \backslash \{k\} : k \leftarrow l \text{ and } k \rightarrow l\}$$
$$D_{k \rightarrow l} = \{l \in S \backslash \{k\} : k \rightarrow l, l \notin D_{k \leftrightarrow l}\}$$
$$D_{k \leftarrow l} = \{l \in S \backslash \{k\} : k \leftarrow l, l \notin D_{k \leftrightarrow l}\}$$
$$D_k = \{l \in S : l \notin D_{k \rightarrow l} \cup D_{k \leftarrow l} \cup D_{k \leftrightarrow l}\}$$

By construction the above sets $D_{k \rightarrow l}, D_{k \leftarrow l}, D_{k \leftrightarrow l}, D_k$ are pairwise-disjoint for every k. Let $\tilde{D}_k := D_{k \rightarrow l} \cup D_{k \leftarrow l} \cup D_{k \leftrightarrow l} \cup D_k$, note that $S = \tilde{D}_k$ for every $k \in S$, hence $\tilde{D}_k$ forms a partition of S. For every $k \in S$ pick any state $l \in \tilde{D}_k$, only one of the following cases must hold true:

- Case 1: $l \in D_{k \rightarrow l}$. If a transition from state k to l occurs at a given timepoint then necessarily we must have $V_k^{(r_1)} = T_l^{(r_2)}$ as we move from k to l. Then $\xi_k^{r_1} \cap \xi_l^{r_2}$ can be similarly evaluated as in previous cases, $\xi_l^{r_2} = \{V_l^{(r_2)} - T_l^{(r_2)} \geq w_{r_2,l}\} = \{V_l^{(r_2)} \geq T_l^{(r_2)} + w_{r_2,l}\} = \{V_l^{(r_2)} \geq V_k^{(r_1)} + w_{r_2,l}\}$ and the intersection gives $\xi_k^{r_1} \cap \xi_l^{r_2} = \{V_l^{(r_2)} \geq V_k^{(r_1)} + w_{r_2,l}\} \cap \{V_k^{(r_1)} \geq T_k^{(r_1)} + w_{r_1,k}\} = \{V_l^{(r_2)} - T_k^{(r_1)} \geq w_{r_2,l} + w_{r_1,k}\}$, now using conditional probability

$$\frac{\mathbb{P}(\xi_k^{r_1} \cap \xi_l^{r_2} | t_k^{(r_1)}, \mathcal{H}(t_k^{(r_1)-}))}{\mathbb{P}(\xi_l^{r_2} | t_k^{(r_1)}, \mathcal{H}(t_k^{(r_1)-}))} = \frac{\mathbb{P}(V_l^{(r_2)} - t_k^{(r_1)} \geq w_{r_1,k} + w_{r_2,l} | \mathcal{H}(t_k^{(r_1)-}))}{\exp\{-w_{r_2,l}\}} \overset{(c)}{=} \frac{\exp\{-(w_{r_1,k} + w_{r_2,l})\}}{\exp\{-w_{r_2,l}\}}$$
$$= \exp\{-w_{r_1,k}\} = \mathbb{P}(\xi_k^{r_1} | t_k^{(r_1)}, \mathcal{H}(t_k^{(r_1)}))$$

(c) Since $T_k^{(r_1)} = t_k^{(r_1)}$ is known and $V_l^{(r_2)}$ is random, by Theorem 2.4.1 the quantity $V_l^{(r_2)} - T_k^{(r_1)}$ is a sojourn time with distribution exp(1).

- Case 2: $l \in D_{k \leftarrow l}$. Under this case we have $V_l^{(r_2)} = T_k^{(r_1)}$, in a similar fashion that in the last case we get $\xi_k^{r_1} \cap \xi_l^{r_2} = \{V_k^{(r_1)} - T_l^{(r_2)} \geq w_{r_1,k} + w_{r_2,l}\}$. So far, to simplify notation, I have avoided

expliciting $T_l^{(r_2)} = t_l^{(r_2)}$ when conditioning, but note that now it is mandatory to explicit that $T_l^{(r_2)} = t_l^{(r_2)}$ when conditioning (otherwise $W_l^{(r_2)}$ would not be a sojourn time).

$$\begin{aligned}\frac{\mathbb{P}(\xi_k^{r_1} \cap \xi_l^{r_2} | \xi_l^{r_2}, t_k^{(r_1)}, t_l^{(r_2)}, \mathcal{H}(t_k^{(r_1)-}))}{\mathbb{P}(\xi_l^{r_2} | t_k^{(r_1)}, t_l^{(r_2)}, \mathcal{H}(t_k^{(r_1)-}))} &= \frac{\mathbb{P}(V_k^{(r_1)} - T_l^{(r_2)} \geq w_{r_1,k} + w_{r_2,l} | t_k^{(r_1)}, t_l^{(r_2)}, \mathcal{H}(t_k^{(r_1)-}))}{\exp\{-w_{r_2,l}\}} \\ &= \frac{\mathbb{P}(V_k^{(r_1)} - t_l^{(r_2)} \geq w_{r_1,k} + w_{r_2,l} | t_k^{(r_1)}, \mathcal{H}(t_k^{(r_1)-}))}{\exp\{-w_{r_2,l}\}} \\ &= \frac{\exp\{-(w_{r_1,k} + w_{r_2,l})\}}{\exp\{-w_{r_2,l}\}} = \exp\{-w_{r_1,k}\} \\ &= \mathbb{P}(\xi_k^{r_1} | t_k^{(r_1)}, \mathcal{H}(t_k^{(r_1)-}))\end{aligned}$$

- Case 3: $l \in D_{k \leftrightarrow l}$. At a given timepoint we can either have $k \to l$ or $k \leftarrow l$, which yields to the proofs of cases 1 and 2.
- Case 4: $l \in D_k$. There's nothing to prove as either $l = k$, which has already been proven, or there's no possible transition between states, so the states are independent by definition.

This ends the proof as the diagram has been proven for all cases. □

These results have crucial consequences as it gives tools to validate MSM. We get the perfect setting to use sojourn times, as residuals to validate MSM, as:

- $W_k^{(r)}$ are always an exponential r.v with expectation 1, and thus the parameter does not depend on any assumption of the model.
- $W_k^{(r)}$ for different $r = 1, \ldots, R$ are independent (thus i.i.d), and so no autocorrelation pattern will be shown if we plot $W_k^{(r)}$ for each r on the x-axis for all states k.

These two points naturally yield to a Central Limit Theorem application. Let $\{\overline{w}_k^{(r)}\}_{1 \leq r \leq R}$ be the observed sojourn times for each state $k \in S$, as each observed sojourn times follows, theoretically, an exp(1) then $S_R(k) = \sum_{r=1}^{R} \overline{w}_k^{(r)} \sim \gamma(\alpha = R, \beta = 1)$ build the standarized residuals

$$Residual_k = \frac{S_R(k) - R}{R} \xrightarrow{\mathbb{P}} \mathcal{N}(0,1), \quad \text{as} \quad R \to \infty, \quad \forall k \in S$$

We now introduce and motivate the definitions of new quantities that are of interest in MSM.

Definition 2.4.1. *(Total time spent in a state) The total time spent in a state $k \in S$ over a time interval $[0, t]$ is denoted as*

$$S_k(t) = \int_0^t I_{\{Z(u)=k\}}(u) du \tag{2.7}$$

Definition 2.4.2. *(Mean sojourn time) The mean sojourn time in state $k \in S$ over a time interval $[0, t]$, for an individual starting in state $s \in S$ at time $t = 0$, and observed fixed covariates vector $\mathbf{X}$, is denoted as*

$$\psi_k(t) = \mathbb{E}[S_k(t) | Z(0) = s, \mathbf{X}] \tag{2.8}$$

These quantities are useful when the states represent different disease conditions and the interest lies in summarizing and comparing the total number of days with specific conditions. A treatment might not be able to prevent infection per se, but it might reduce the infection time, thus the cumulative mean time spent on an "infection" state would be useful to test the hypothesis of a positive effect of treatment.

Other contexts in which these quantities are useful is utility-based analysis. For instance, in health economic analysis it is frequent to calculate the incurred cost of a policy based on hospitalization days or discharge, mean sojourn times for hospitalization or discharge states are estimates of these quantities.

Proposition 2.4.1. *If $\psi_k(t)$ denotes the expected total sojourn time in a state $k \in S$ over a time interval $[0, t]$, then*

$$\psi_k(t) = \int_0^t \mathbb{P}(Z(u) = k | Z(0) = s, \mathbf{X}) du \tag{2.9}$$

Proof. Let $t > 0$ be fixed. Note that $\mathbb{E}\left[|S_k(t)|\right] = \mathbb{E}[S_k(t)]$ and $S_k(t) = \int_0^t I_{\{Z(u)=k\}}(u)du \leq \int_0^t du = t < \infty$, hence $\mathbb{E}[S_k(t)] < \infty$ and

$$\begin{aligned}\psi_k(t) = \mathbb{E}\left[S_k(t)|Z(0)=s,\mathbf{X}\right] &= \mathbb{E}\left[\int_0^t I_{\{Z(u)=k\}}(u)du\Big|Z(0)=s,\mathbf{X}\right] = \int_0^t \mathbb{E}\left[I_{\{Z(u)=k\}}(u)|Z(0)=s,\mathbf{X}\right]du \\ &\overset{(a)}{=} \int_0^t \mathbb{P}(Z(u)=k|Z(0)=s,\mathbf{X}) + 0\cdot(1-\mathbb{P}(Z(u)=k|Z(0)=s,\mathbf{X}))du \\ &= \int_0^t \mathbb{P}(Z(u)=k|Z(0)=s,\mathbf{X})du\end{aligned}$$

(a) For a fixed $u > 0$, $I_{\{Z(u)=k\}}(u)$ is a Bernoulli random variable with probability $\mathbb{P}(Z(u)=k|Z(0)=s,\mathbf{X})$. □

To end with this section, we define the survival function for a MSM, which is the time until entering one of the multiple absorbing states.

Definition 2.4.3. *(Survival function) Let $A \subset S$ be the set of all absorbing states and the survival time as $T = \inf_{a\in A}\{t \in \mathbb{R}^+ : Z(t) = a\}$ then the survival function is defined as*

$$S(t) = \mathbb{P}(T \geq t)$$

2.5 Transition probabilities in time-homogeneous Markov process

In this section we will explore how this transition probabilities can be computed for a time-homogeneous Markov process, this section is built from scratch, as the primary reference [8] treats a more general case with minor detail on time-homogeneous Markov process.

Transition probabilities are one of the most powerful tools of a MSM, as it gives the probability that a subject transitions from one state into another given its profile through covariates $\mathbf{X}$. All other relevant tools, that might be useful to test hypothesis, are derived from transition probabilities in one way or another: survival times, sojourn times, mean sojourn times, and so on. Fixed-time covariates are assumed $\mathbf{X}$, and either a multiplicative or additive form can be assumed.

Under time-homogeneous Markov process assumption, transition intensity functions are of the form $\lambda_{kl}(t|\mathcal{H}(t^-)) = \lambda_{kl|\mathbf{X}}$ from state $k \to l$. We will denote $Q_{|\mathbf{X}}$ as the transition intensity matrix without assuming a specific functional form of the transitions.

$$Q_{|\mathbf{X}} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{K1} & \lambda_{K2} & \dots & \lambda_{KK} \end{pmatrix}_{|\mathbf{X}} \tag{2.10}$$

For (2.10) to be a transition intensity matrix properties of definition 1.2.12 must hold.

As in almost every case, transition intensities will not be known and an estimation of these will be needed. The estimation of these transition intensity functions $\hat{\lambda}_{kl|\mathbf{X}}$, via either parametric or semi-parametric regression, must satisfy first and second properties of definition 1.2.12, otherwise no guarantees that $\hat{Q}_{|\mathbf{X}}$ is a transition intensity matrix and so the computations of the transition probability matrix will be incorrect. If one assumes an additive form then the estimations of $\hat{\lambda}_{kl|\mathbf{X}}$ might not be positive per se, and some restrictions should be imposed in the estimation procedure to avoid this case.

Taking into account all the previous points, we rewrite the transition intensity matrix as:

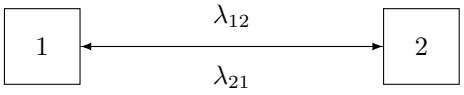
$$Q_{|\mathbf{X}} = \begin{pmatrix} -\sum_{l\in S\setminus\{1\}}\lambda_{1l} & \lambda_{12} & \dots & \lambda_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{K1} & \lambda_{K2} & \dots & -\sum_{l\in S\setminus\{K\}}\lambda_{Kl} \end{pmatrix}_{|\mathbf{X}} \tag{2.11}$$

Since $Q_{|\mathbf{X}}$ is a transition intensity matrix using Theorem 1.2.1, $P(t)_{|\mathbf{X}} = e^{tQ_{|\mathbf{X}}}$ is a stochastic matrix, and by Theorem 1.2.2: $P(t)_{|\mathbf{X}} = e^{tQ_{|\mathbf{X}}}$ is the transition probability matrix as it satisfies both the Forward and Backward Kolmogorov differential equations.

This section gives an idea on the natural estimation process for MSM:

1. Set the regressions for the transition intensity functions $\lambda_{kl|\mathbf{X}}$
2. Estimate all the regressions for the transition intensity functions, with this we get the estimation of the transition intensity matrix $\hat{Q}_{|\mathbf{X}}$
3. Use matrix exponential to compute $e^{t\hat{Q}_{|\mathbf{X}}}$, which is finally the transition probability matrix at time t.

Example) Assume a time-homogeneous Markov bidirectional MSM with $S = \{1, 2\}$. For each subject two covariates have been observed $X_1 = Sex$ and $X_2 = Wage$ (euros). The corresponding diagram is



In this case we just need two regressions, the regressions are set of the form

$$\lambda_{21|\mathbf{X}} = \lambda_{21,0} \exp\{\beta_{21,1} X_1 + \beta_{21,2} \log(X_2)\}$$

$$\lambda_{12|\mathbf{X}} = \lambda_{12,0} \exp\{\beta_{12,1} X_2\}$$

And for some estimates of $\beta_{12,1}, \beta_{21,1}, \beta_{21,2}, \lambda_{12,0}, \lambda_{21,0}$ we build the estimated transition intensity matrix as

$$\hat{Q} = \begin{pmatrix} -\hat{\lambda}_{12,0} \exp\{\hat{\beta}_{12,1} X_2\} & \hat{\lambda}_{12,0} \exp\{\hat{\beta}_{12,1} X_2\} \\ \hat{\lambda}_{21,0} \exp\{\hat{\beta}_{21,1} X_1 + \hat{\beta}_{21,2} \log(X_2)\} & -\hat{\lambda}_{21,0} \exp\{\hat{\beta}_{21,1} X_1 + \hat{\beta}_{21,2} \log(X_2)\} \end{pmatrix}$$

The transition probability matrix at time t could be computed by first diagonalization of $t\hat{Q}$ and then exponentiating the diagonalized matrix.

Chapter 3

Inference on Markov multi-state models

3.1 Likelihood function construction

This section is based on section 2.2.1 from [10], I detail every step and extend when required. The likelihood function is used to; compute estimators, either parametric or semiparametric, for the regression of transition intensities[1]; compute information criterion (AIC, BIC, Schwarz); statistical hypothesis testing, and so on. Its form differs broadly depending on:

- Censoring pattern of states. In this section and throughout we will focus on right-censoring and so other types of censoring: random censoring, left censoring and interval-censoring, will not be considered.
- Existence of time dependent covariates in the regression of the transition intensities. I will not consider time-varying covariates for the regressions, as that is a more advanced topic beyond the scope of thesis' objectives.

Despite studying a concrete type of MSM (homogeneous Markov MSM with, at most, a right censoring pattern), the derivation of the likelihood will not have any specification towards which stochastic process is used (Markov, semi-markov, non-Markov). Our aim is to construct a likelihood function such that it contains the proper information to estimate transition intensity functions. In this section we make some assumptions, to ease the derivation of the likelihood, as the primary goal of this section is to get an idea of the likelihood construction of such a process without diving into more advanced topics.

- Right-censoring based on a fixed administrative censoring time.
- No time-dependent covariates
- Subject states are known only at intermittent observation times, visit times are fixed an homogeneous for all subjects. This assumption is too strong when there is existence of lost to follow-up subjects, or when there are subjects that skip assessment visits.
- Visit times are fixed based on study protocol. In some studies this assumption might not hold, for instance in disease registries clinic visits may be set by physicians sequentially such that next visit is scheduled based on previous condition of patient, this is known as *conditionally dependent visit process.*
- Subjects are perfectly classified, there exists a rule to classify subjects such that its missclassification error is 0. This assumption is not too strong, in some clinical studies a patient is classified in a state if multiple measures agree on classification, which drastically reduces the missclassification error.

Assume we are in a study that observes an individual over a time interval $[0, C^A]$, where C^A is a pre-study fixed censoring administrative time, for instance end of study date. Actually, we do not observe subjects on a study at all time points, but at a finite set of it, say $0 = u_0 < u_1 < \cdots < u_M = C^A$.

[1] Note that we just need to estimate transition intensity functions to get all information about: transition probabilities, sojourn times, survival times, etc.

Because the likelihood function must contain all the information about the multi-state process, and a multi-state process is defined via $Q_{|\mathbf{X}}$ our data must properly represent $Q_{|\mathbf{X}}$, then our sample for a given time visit u_m will be a matrix. Such a matrix will empirically emulate the transition intensity matrix $Q_{|\mathbf{X}}$, hence each element of the sample matrix will be the number of transitions observed over a time interval between visits. To fix ideas, a picture of what the realization of a multi-state process looks like is presented.

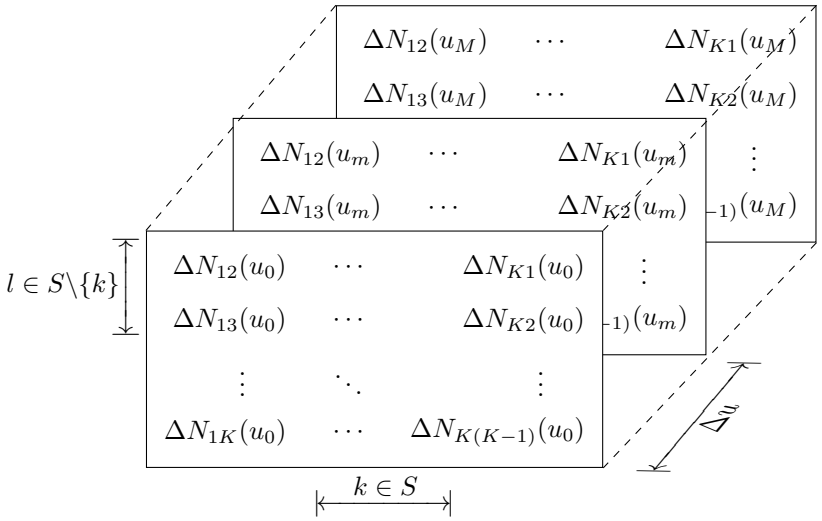


Figure 3.1: Conceptualization of a realization of a multi-state process. Height and length represent states, whereas width represents time. Each element represents the number of transitions from column state to row state over time interval $[u_{m-1}, u_m)$.

Source: Own elaboration

For each visit, information about the number of transitions from one state into another over a time interval $[u_{m-1}, u_m)$ is collected. The realization of a multi-state process is a hypermatrix as figure 3.1 shows, note that we just have to estimate $K(K-1)$ transition intensity functions and so we only need data on $K(K-1)$ observed transitions, as the diagonal entries are fixed. A sample of the hypermatrix at a given time u_m is

$$\Delta N(u_m) = (\Delta N_1(u_m)', \ldots, \Delta N_K(u_m)')' = \begin{pmatrix} \Delta N_{12}(u_m) & \ldots & \Delta N_{K1}(u_m) \\ \Delta N_{13}(u_m) & \ldots & \Delta N_{K2}(u_m) \\ \vdots & \ddots & \vdots \\ \Delta N_{1K}(u_m) & \ldots & \Delta N_{K(K-1)}(u_m) \end{pmatrix} \tag{3.1}$$

and denote $\mathcal{H}(u_m) = \{Z(u_r), 0 \leq r \leq M\}$[2]. A realization of the process can then be expressed as

$$\mathcal{L} = \prod_{m=1}^{M} \mathbb{P}(\Delta N(u_m) | \mathcal{H}(u_{m-1})) \tag{3.2}$$

To simplify notation, an abuse of notation has been made when avoiding the set to which the probability is taken. As $M \to \infty$ and $\Delta u_r \to 0$, the probability of two or more transitions ocurring within $[u_{m-1}, u_m)$ vanishes to 0. With M large enough, only one transition can occur, and so we can consider the different contributions depending on state occupied at u_{m-1}.

$$\begin{aligned} \mathcal{L} = \prod_{m=1}^{M} \mathbb{P}(\Delta N(u_m) | \mathcal{H}(u_{m-1})) &= \prod_{m=1}^{M} \prod_{k \in S} \mathbb{P}(\Delta N_k(u_m) | \mathcal{H}(u_{m-1}))^{Y_k(u_{m-1})} \\ &= \prod_{m=1}^{M} \prod_{k \in S} \prod_{l \in S\backslash\{k\}} \mathbb{P}(\Delta N_{kl}(u_m) | \mathcal{H}(u_{m-1}))^{Y_k(u_{m-1})} \end{aligned}$$

[2]As stated in previous sections, this filtration is equivalent to $\mathcal{H}(u_m) = \{N(u_r), 1 < r \leq M, Z(0)\}$

Remember that $Y_k(u_m) = I_{\{Z(u_m)=k\}}(u_m)$ is the indicator that the process is in state k at time u_m. As $M \to \infty$, $\Delta N_{kl}(u_m)$ is a bernoulli random variable, $\Delta N_{kl}(u_m) \overset{a}{\sim} Bern(p_{kl}(u_m))$ where

$$p_{kl}(u_m) = \mathbb{P}\left(\{\exists l \in S\backslash\{k\} : \Delta N_{kl}(u_m) = 1\}|\mathcal{H}(u_{m-1})\right) \tag{3.3}$$

$$q_{k\cdot}(u_m) = \mathbb{P}\left(\{\forall l \in S\backslash\{k\} : \Delta N_{kl}(u_m) = 0\}|\mathcal{H}(u_{m-1})\right) = 1 - \sum_{l\in S\backslash\{k\}} p_{kl}(u_m) \tag{3.4}$$

The probability mass function of $\Delta N_{kl}(u_m)$ is:

$$\mathbb{P}(\Delta N_{kl}(u_m) = x|\mathcal{H}(u_{m-1})) = p_{kl}(u_m)^x q_{k\cdot}(u_m)^{1-x}, x = 0, 1 \tag{3.5}$$

Using the probability mass function (3.5):

$$\mathcal{L} = \prod_{m=1}^{M}\prod_{k\in S}\prod_{l\in S\backslash\{k\}}\left[p_{kl}(u_m)^{\Delta N_{kl}(u_m)} q_{k\cdot}(u_m)^{1-\Delta N_{k\cdot}(u_m)}\right]^{Y_k(u_{m-1})} \tag{3.6}$$

$$= \prod_{m=1}^{M}\prod_{k\in S}\left[q_{k\cdot}(u_m)^{1-\Delta N_{k\cdot}(u_m)} \prod_{l\in S\backslash\{k\}} p_{kl}(u_m)^{\Delta N_{kl}(u_m)}\right]^{Y_k(u_{m-1})} \tag{3.7}$$

The core of a MSM are transition intensity functions, it is natural then that the likelihood is expressed using those. By first using proposition 2.3.1 and definition 2.2.2 after, we can rewrite (3.6) in terms of transition intensity functions:

$$\begin{aligned}\mathcal{L} = \prod_{m=1}^{M}\prod_{k\in S}\Bigg[&\left\{1 - Y_k(u_{m-1})\sum_{l\in S\backslash\{k\}}\lambda_{kl}(u_m|\mathcal{H}(u_{m-1}))\Delta u_m + o(\Delta u_m)\right\}^{1-\Delta N_{k\cdot}(u_m)} \\ &\prod_{l\in S\backslash\{k\}}\left(Y_k(u_{m-1})\lambda_{kl}(u_m|\mathcal{H}(u_{m-1}))\Delta u_m + o(\Delta u_m)\right)^{\Delta N_{kl}(u_m)}\Bigg]^{Y_k(u_{m-1})}\end{aligned} \tag{3.8}$$

We can take the product in (3.8) in the order $\prod_{k\in S}\prod_{l\in S\backslash\{k\}}\prod_{m=1}^{M}$ and divide it by $\prod_{k\in S}\prod_{l\in S\backslash\{k\}}\prod_{m=1}^{M}(\Delta u_m)^{\Delta N_{kl}(u_m)}$ and then take the limit as $M \to \infty$. For a given pair of indices (k, l), the terms $o(\Delta u_m)/\Delta u_m \to 0$ as $R \to \infty$, the terms $Y_k(\cdot)\lambda_{kl}(\cdot|\mathcal{H}(\cdot))$ remain for the times at which a transition occurs from $k \to l$, since it is when $dN_{kl}(u) = 1$. Those intervals not containing transitions will have contributions represented outside the curly bracket in (3.8), Proposition 1.3.2 gives

$$\exp\left(-\int_0^{C^A} Y_k(s)\lambda_{kl}(s|\mathcal{H}(s^-))ds\right)$$

And the likelihood (3.8) as $M \to \infty$ becomes:

$$\mathcal{L} = \prod_{k\in S}\prod_{l\in S\backslash\{k\}}\left[\prod_{t_j\in D_{kl}}\lambda_{kl}(t_j|\mathcal{H}(t_j^-))\exp\left(-\int_0^{C^A} Y_k(s)\lambda_{kl}(s|\mathcal{H}(s^-))ds\right)\right] \tag{3.9}$$

Where D_{kl} denotes the set of $k \to l$ transitions times over $[0, C^A]$.

3.2 Maximum likelihood estimation for parametric regression

As we only focus on time-homogeneous Markov MSM with multiplicative form, all regressions are of the type

$$\lambda_{kl}(t|\mathcal{H}(t^-), \mathbf{X}) = \lambda_{kl,0} \exp\{\beta'_{kl}\mathbf{X}\} = \exp\Big\{\theta'_{kl}\tilde{\mathbf{X}}\Big\} \tag{3.10}$$

Where $\theta_{kl} = (\lambda_{kl,0}, \beta'_{kl})$ and $\tilde{\mathbf{X}} = (1, \mathbf{X})$, the regressions are fully parametric as $\lambda_{kl,0}$ does not depend on time. We stick to the case where observed transitions times have, at most, a right-censoring pattern. The estimators $\hat{\theta}_{kl}$ are given by maximum likelihood, only in a few cases these estimators can be computed analytically, otherwise non-linear optimization methods are used. Some properties of these estimators are:

- $\sqrt{n}(\hat{\theta}_{kl} - \theta_{kl}) \overset{a}{\sim} \mathcal{N}_p(\mathbf{0}_{p\times 1}, \hat{I}_{kl}(\hat{\theta}_{kl})^{-1})$,where $\hat{I}_{kl}(\hat{\theta}_{kl})$ is the estimated Fisher information matrix for the $k \to l$ transition.
- Confidence intervals and Wald tests are based on the asymptotic normal distribution of $\hat{\theta}_{kl}$.
- $Var(\hat{\theta}_{kl})$ is computed by the multivariate Delta Method.

For our case, $\hat{\theta}_{kl}$ can be computed analytically and so a global maximum is always guaranteed when using iterative numerical methods. `R` software to fit MSM often requires an initial transition intensity matrix $Q^{(0)}_{|\mathbf{X}}$ so as to initialize the iterative algorithm to compute MLE. Depending on that initial condition we can either get: global maximum, local maximum, non convergence of the iterative method, for that reason, it is recommended to use different initial conditions to analyze convergence upon a global maximum. In this thesis, we will just apply time-homogenous Markov MSM to real data, and an analysis of such convergence of the iterative method based on different $Q^{(0)}_{|\mathbf{X}}$ will not be necessary as we will always attain a global maximum.

Chapter 4

Application to real dataset

4.1 Generic data structure for multi-state models and R packages

Multi-state models can be fitted using software that has been developed for survival analysis, though many new specific packages to fit MSM in R software have been developed. Parametric models can be tackled with `msm` package, this includes: time-homogeneous Markov models, non-homogeneous Markov models, hidden Markov models, we will stick to the time-homogeneous Markov models application. Methods of estimation and model assessment are available in this package, and `msm` will also fit Markov models to continuously observed data. Another useful package is `mstate`, this handles right-censored data and allows for nonparametric estimation (we will stick to the parametric estimation case), and fits semiparametric Cox models for intensity functions (i.e non-homogeneous Markov with multiplicative functional form), it also provides for good visualization tools that are not available in `msm` package.

The data for each study can differ broadly, as each study has its special characteristics which, eventually, reflects on the data structure used. Furthermore, the data for each study can be recorded in a variety of forms. A common feature of MSM´s datasets is a longitudinal type dataset. Several subjects will be observed over times, and so in any MSM we will have a repeated measures scheme. We consider a generic dataset structure for a study that does not incorporate time-varying covariates.

For a study with n subjects, 3 states, with directions $\{1 \to 2,\ 1 \to 3,\ 2 \to 3\}$, and p fixed covariates X_p, one way to record the data of a multi-state process is[1]

```
> MSM_data
```

	id	start	stop	from	to	status	X1	X2	...	Xp	gtime
1	id_1	0	68	1	2	1	2.3	1		4	68
2	id_1	0	68	1	3	0	2.3	1		4	68
3	id_1	68	749	2	3	1	2.3	1		4	710
4	.	0	.	.	.	.	.	.		.	.
5	.	.	.	.	.	.	.	.		.	.
6	.	.	.	.	.	.	.	.		.	.
7	id_n	0	39	1	2	1	3.4	0		6	39
8	id_n	0	39	1	3	0	3.4	0		6	39
9	id_n	39	100	2	3	1	3.4	0		6	61

The variable `id` is the label associated with each individual[2], and it is unique, we have denoted this variable with an index i in the theoretic chapters, though we have avoided sometimes this subindex to simplify notation. As each observation is within a time interval, variables `start` and `stop` contain the left and right endpoints of these time intervals, respectively. `Status` variable indicates whether or not this intervals ends with a transition out of the state recorded in the `from` column to the `to` column. `gtime` variable is the observed sojourn time for the state in the `from` column. This way of recording data

[1] No censoring pattern has been considered. For complex censoring patterns, data structure can vary broadly. Otherwise for usual censoring patterns (left, right, interval), it suffices to add a variable `censor` that takes value 1 if observation is censored and 0 otherwise.

[2] In `msm` package, if no id is found, all observations are assumed to be from the same individual.

is way more inefficient than a counting process structure one, as subjects can be at risk for transitions from a given state into two or more states at a given time, there may be multiple rows in the dataframe corresponding to each at-risk period for an individual.

For the same data-type, consider a dataset that records information of the multi-state process in a counting process structure.

```
> MSM_data

    id start stop from to  X1 X2 ... Xp
1  id1     0   68    1  2 2.3  1      4
2  id1    68  749    2  3 2.3  1      4
3    .     .    .    .  .   .  .      .
4    .     .    .    .  .   .  .      .
5    .     .    .    .  .   .  .      .
6 id_n     0   39    1  2 3.4  0      6
7 id_n    39  100    2  3 3.4  0      6
```

With a counting process structure, there is no need to represent each at-risk period for an individual, we just capture transitions over assessments, in this way we get the same information as the last data-type structure but in a more compact manner.

Counting process structure is based on counting process notation that was introduced in previous thereotic sections of Multi-state model theory, thus this data structure provides a neat link between developed theory and real data. Furthermore, not only counting process structure reduces the amount of rows needed to represent information, but also the amount of variables needed to represent such information, it is only necessary to include either a `from` variable or a `to` one, but we include both here for clarity and consistency with how more complex dataframes are constructed.[3] `Status` variable is not required with this data structure, as we only represent the observations with a given transition and avoid registering all possible at-risk transitions, thus `Status` variable becomes obsolete in this data structure. Finally, `gtime` can be directly derived from `stop` − `start` for each observation, because we are not registering all at-risk transitions.

In the dataset we will be using for applications, we will stick to the counting process structure, as it is a more convenient and natural structure to fit models with packages `msm` and `mstate`.

[3] As an extra: time-dependent covariates dataframes need both variables `from` and `to` to keep data consistent.

4.2 Targeted disease control of individual Meerkats (*Suricata suricatta*) against Tuberculosis

4.2.1 Motivation and objectives

There is evidence, that in a population, some individuals are more likely to spread disease while others are more susceptible to be infected. In human population, the vaccination strategy aims to vaccinate as much individuals as possible so as to erradicate the spread of a disease, but that strategy might fail when it comes to vaccination in wildlife animals. To erradicate a disease, it is as important to vaccinate wildlife animals as to vaccinate human population, because some diseases are well-known to be zoonotic-diseases. [4] An example of such is tuberculosis disease.

Most of the vaccination strategies assume that the disease spreads out in an homogeneous and random way [1]. However, if that assumption does not hold, then vaccinating as much individuals as possible is not the optimal strategy based on: infection rates, time, economic cost of vaccines, in such case we would not be optimally allocating limited resources. Even if the assumption did hold, it is hard to erradicate the disease by vaccinating all animals as: it is hard to estimate population numbers, efficiency of delivery, unintentional repeated doses[4].

To this end, two trait-vaccination strategies will be empirically tested, as Patterson et al. [14] does, but from a completely different methodological point of view. The original paper claims that their analysis contributes to the understandings of: social network structure, heterogeneity within the system, and knowledge of the infection dynamics within the system. Their analysis is based on a Cox regression for the endpoint time to death. Such an analysis does not allow to extract information on the dynamics of the disease, as opposed to a MSM methodology which would enrich the analysis introducing multiple endpoints of interest as: time to death, time to test positive for tuberculosis, time to euthanasia. For these reasons, we consider that a MSM approach would be more suitable for the posed objectives.

1. **High contact** (HC): individuals with typically the highest rates of social contact, dominant individuals (either females or males) in meerkat social hierarchy.
2. **High-susceptibility** (HS): individuals who by their social status have previously been shown to be the most susceptible to be infected by tuberculosis, subordinate animals.

Two vaccination strategies are defined based on these social hierarchy. One vaccination strategy will consist on vaccinating only high-contact meerkats of a social group. On the other hand, second strategy will only vaccinate high-susceptibility individuals (young subordinates). A control group will be used too. The best vaccination strategy will be the one with the highest survival times and less infection rates.

Some of the objectives of this application are:

1. To find an optimal allocation strategy, based on significant traits related to infection of Tuberculosis, for vaccination of wildlife individual Meerkats, so as to minimize the rate of infection of Tuberculosis disease in such population.
2. Study the dynamics of tuberculosis disease in a population of wildlife Meerkats.
3. To show the full process of building a time-homogeneous Markov MSM to answer all previous questions.

4.2.2 Design of the experiment

This experiment was ran by [14], we explain how the data was generated as it is relevant to set up the MSM. The study was carried out at Kuruman River reserve in South Africa's Nothern Cape. All animals included in the study were part of the Kalahari Meerkat project's long-term study of a free-ranging meerkat population. All individuals were identifiable by pre-existing dye marks. The study was ran from 1 September 2014 to 30 September 2016.

Treatment (vaccination) sets (strategies), carried out at first sampling event, were defined as:

[4] Zoonotic-diseases are the diseases or infections which are naturally transmitted between animals and humans.

(1) high-susceptibility: only dominant individuals (female and male) were vaccinated.

(2) high-contact: all subordinate animals were vaccinated. No dominant individuals were vaccinated.

(3) control: no individual of the group was vaccinated.

Nine social groups of meerkats were included in the study. Note that this groups are not meant to be fixed over time, there were nine social groups at start of the study but those groups evolved over the course of it. Nevertheless, treatment allocation was based on those nine initial groups. Furthermore, groups were ranked into three tiers: large groups (>20 individuals), medium (10-20), and small (<10). The treatment allocation strategy was as follows:

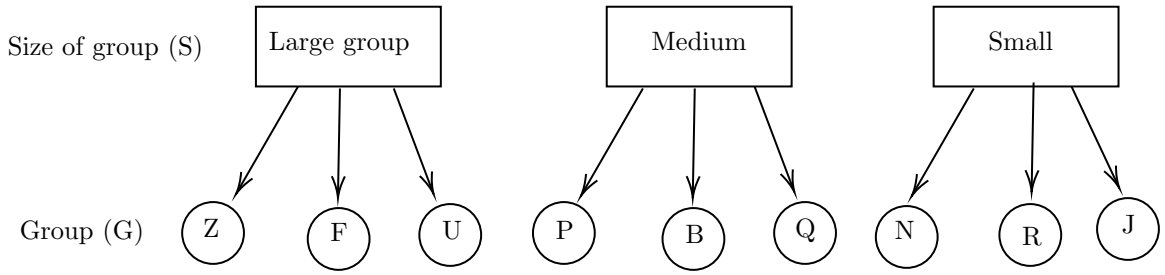


Figure 4.1: Generation of the random allocation of treatments by initial group.

Source: Own elaboration.

Factor G is nested to S, for each G(S) a randomly allocated treatment was assigned, and so each treatment is well-represented by all group-sizes, though treatment clearly depends on G, as T is nested to G (i.e T(G)). As T is nested to G, it will not be possible to separate group effect on treatment.

Visit schemes are periodic based on predefined scheduled, so visits schedules guarantee non-informative censoring and non-intermettent visit data, which are good properties for a MSM to have. There was five equal time blocs, and so, at maximum, five measures of each individual have been taken, with time points:

Block 1: September-November 2014

Block 2: December 2014-February 2015

Block 3: July-October 2015

Block 4: January-March 2016

Block 5: July-September 2016

4.2.3 Dataset

The dataset was extracted from https://figshare.com/authors/_/2886764. Not all variables from the original dataset were used (39 variables), as most of them were time-dependent covariates. Time-dependent covariates is beyond the scope of this thesis, and so all time-dependent covariates were removed from the analysis, treating a time-dependent covariates case is not the primary objective of the thesis. Dataset´s name is "*Life time Survival*", which contains data of 274 individual wild meerkats with samples over time, the data measures the time until: all-cause mortality, testing positive for Tuberculosis for the first time. A table with a description of all variables extracted from the original dataset, that were either included in the analysis or used as metadata, is presented.

Variable's name	Variable type	Description
Sample.No	Integer	Sample number
id	Character	Unique identifier of Meerkat. It is composed by "VXXYY", where "XX" is its sample group and "YY" its number assigned among all Meerkats in given sample group.
Sample Date	Date	Date of sample extraction.
Sex	Factor	Takes value 1 if Meerkat biological sex is a female and 2 otherwise.
euth	Factor	Takes value 1 if Meerkat was euthanized and 0 otherwise.
dth	Factor	Takes value 1 if Meerkat is considered to have died and 0 otherwise. Meerkat is considered to have died if it was not observed over a period of three months.
time	Integer	Difference, in days, between stop and start dates
Val&Pos	Factor	Takes value 1 if positive by tuberculosis and 0 otherwise
event	Factor (DERIVED)	Derived variable. Takes values 1 (Tuberculosis-free), 2 (Active infection of tuberculosis), 3 (dead by euthanasia), 4 (dead by any cause but euthanasia).
tx_grp	Factor	Takes values: 1 if treatment is high-susceptibility (HS), 2 for high-contact (HC), 3 to control group.
start	Integer	Day at which the sample was observed
stop	Integer	Last day the sample was observed

Table 4.1: Description of the variables included in the analysis

Source: Own elaboration.

Variable `event` has been derived from other variables, in the original dataset, to create states. Before delve into the procedure used to generate such variable, it is relevant to deepen into how variables used to derived `event` were measured:

1. `dth` : an individual was considered to have died if it was not observed for a period of three months.

A censoring pattern emerges from the nature of this data, all transitions from any state to death state should be interval-censored, though we will simplify the case and not consider censoring as the original dataset does not provide insight into other censoring patterns but just interval-censoring.

2. `euth` : individuals could be captured, out of predefined time windows, for other reasons such as euthanasia, in such case an opportunistic sample were taken at that point, although visit not being into the predefined time blocks.

3. `Val&Pos` : at each sampling event the individual was manually caught and sampled under general anaesthesia, multiple measures were collected: blood sample, DPP VetTB[5], IPRA assay[6], tracheal lavage for mycrobacterial culture. An individual was classified as test positive (i.e `Val&Pos` = 1) if it fulfilled one or more of the following criteria:

   - IPRA result greater than 0.038.
   - DPP VetTB greater than 5.0 Reflective Light Units (RLU).
   - Confirmed positive by PCR from tracheal lavage sample.

   Variables used to classify for Tuberculosis can be found on the original dataset. As the original study did not consider classifying individuals on states and at least one of the criterias sufficed to classify as positive: existence of missclassification error could arise in this study depending on false-negative and false-positive rates of tests. For our proposed methodology, a better strategy to classify individuals as positive would have been if multiple criteria (at least 2) agreed on Tuberculosis status classification.

The next diagram shows the logic used to derive variable `events` that define the states.

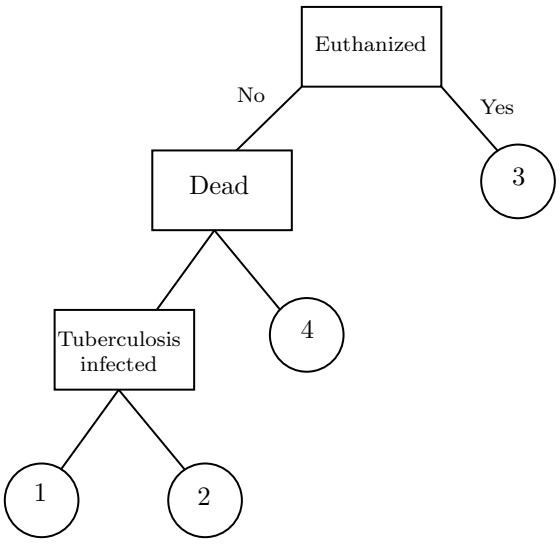


Figure 4.2: Derivation of states through variables: `euth` (Euthanized), `dth` (Dead), `Val&Pos` (Tuberculosis infected). Values for `event`: (1)- Tuberculosis free; (2)- Tuberculosis infected; (3)- Euthanized; (4)- Dead due to other cause but euthanasia.

Source: Own elaboration.

States are defined as: (1)- Tuberculosis-free; (2)- Tuberculosis infected; (3)- Euthanized; (4)- Dead due to other cause but euthanasia. Variables `dth` and `Val&Pos` contained NA values, in such cases a 99 value for `event` was assigned[7]. If, for a given observation, no previous variables had NA values then 4.2 was used to generate states. The hierarchy of the tree is based on how original dataset codified information. If an individual had been euthanized then `euth` = 1 and `dth` = 1, but the actual cause of death was euthanasia and so `event` = 3 despite `euth` = `dth` = 1 at the same time. On the other hand,

[5]Dual Path Platform (DPP) VetTB assay, a lateral-flow type test for detecting antibodies to Tuberculosis.
[6]Analysis of IP-10 release assay, to test for Tuberculosis status.
[7]Imputation methods such as LCFO (Last Carried Forward Observation) could be used, but note that using imputation methods on `event` should require an extensive analysis as it could increase the missclassification error thus giving biased results.

if `euth` = 0 and `dth` = 1, then the cause of death is other than euthanasia, hence `event` = 4. Finally, if `euth` = `dth` = 0 individual is alive and state depends on either testing positive or not, if the individual tests positive then `Val&Pos` = 1 and so `event` = 2, otherwise `event` = 1.

Definition of states was made based on objectives of the analysis. Note that the states: (3)- euthanized, (4)- death by other cause but euthanasia, define a competing risks situation, as an individual can experience just one of the states throuhgout its life cycle, and two states are absorbing states. The decision to split death state into the former ones is to get a deeper insight into the causes of death, so as to study the different patterns of death cause by treatment, or other statistically significant covariates that might interfere in the output. However, the original paper does not split death based on cause of euthanasia, then if an individual dies from euthanasia the observation is treated as interval-censored when actually its not censored at all (deaths from euthanasia are non-censored). With our metholodogy given that death is interval-censored but euthanasia is not even censored, we can distinguish the censoring patterns: transitions to euthanasia state won't have any censoring pattern, transitions to "dead due to other cause but euthanasia" will be interval-censored. Unfortunately, we are not considering any censoring pattern, as the proper one should be interval-censored but that is beyond the scope of the thesis.

4.2.4 Data cleaning and descriptive statistics

Data has been pre-processed via `R` software, the code used in this section is attached in Appendix 4.2.6. Although data comes from a peer-reviewed journal, the raw dataset is not conveniently structured for the methodology we propose and several consistency checks must be performed to ensure the methodology is correct. First, variable `event` was created with the logic shown in Figure 4.2 through self-programmed function `classify(death, positive, euth)`, which receives the triplet `(death, positive, euth)` and returns a value: 1, 2, 3, 4 or 99. The first problem in the data cleaning process arises when classifying subjects in states at every time point, if `event` = 99 then either: `death = NA`, `euth = NA`, `positive = NA`, hence we don't have enough information to classify the subject and its state is unknown at that timepoint.

Next step is to derive variable `Time.blocks` through self-programmed function `timeblock(x)` that receives a *Sample.Date* and returns the time block the sample belongs to, the function can return either one of these values: Block 1, Block 2, Block 3, Block 4, Block 5, Out of time windows. Turns out that there are quite a few of observations which are out of time windows predefined in the design. We will tackle all these problems laters and a justification of the decisions will be made. Only two covariates will be considered for the analysis: sex (`sex`), treatment (`tx_grp`), the reason is that these are the only fixed-time covariates we can include in the study as the original dataset included much more covariates, but unfortunately these are time-varying covariates which is a more advanced case we are not considering.

A consistency check has been done for the covariates `sex` and `tx_grp` to check that those are actually fixed-time covariates, and that individuals do not experience changes on those throughout the study. Again, a self-programmed function `consistency_fix_time(x)` has been used for this check, as no function is available in `msm` package to carry out this check[8]. After running the former function, we conclude all variables are fixed-time as a `NULL` value is returned for all variables when running the function. Since all variables are factors, we present the descriptive summary (count and %) of variables: `event`, `sex`, `tg_grp`, by variable `Time.block`, to observe at each sample point what is the distribution of each factor. For the sake of transparency and clarity, we will first present the descriptive summary for the raw dataset, which is the dataset for which variables `Time.block` and `event` have been created, and later on we will present the same summary for the clean dataset (after some actions taken on the raw dataset that we will explain later).

[8]The function `statetable.msm` from `msm` might do the job, as it creates a crosstable with the baseline values of the covariates and its values at the last observation for each subject, but this might be too risky to use as a subject might not experience a change on *sex* covariate from baseline to last visit, but it might, due to some data-entry error, for instance from baseline to second visit which would not be noticed by this function.

Descriptive statistics

Based on raw dataset LifeSurvivalDataset

Variable[1]	Block 1 N = 83[1]	Block 2 N = 128[1]	Block 3 N = 183[1]	Block 4 N = 81[1]	Block 5 N = 77[1]	Out of time windows N = 101[1]
sex						
F	35 (42%)	48 (38%)	70 (38%)	37 (46%)	36 (47%)	40 (40%)
M	48 (58%)	80 (62%)	109 (60%)	43 (53%)	41 (53%)	56 (55%)
			4 (2.2%)	1 (1.2%)		5 (5.0%)
tx_grp						
1	27 (33%)	44 (34%)	51 (28%)	28 (35%)	13 (17%)	21 (21%)
2	29 (35%)	31 (24%)	57 (31%)	28 (35%)	26 (34%)	38 (38%)
3	27 (33%)	53 (41%)	75 (41%)	25 (31%)	38 (49%)	42 (42%)
event						
1	43 (52%)	70 (55%)	77 (42%)	33 (41%)	22 (29%)	33 (33%)
2	20 (24%)	10 (7.8%)	28 (15%)	12 (15%)	17 (22%)	11 (11%)
3	3 (3.6%)	1 (0.8%)	4 (2.2%)	2 (2.5%)	12 (16%)	4 (4.0%)
4	9 (11%)	17 (13%)	73 (40%)	32 (40%)	3 (3.9%)	22 (22%)
99	8 (9.6%)	30 (23%)	1 (0.5%)	2 (2.5%)	23 (30%)	31 (31%)

[1] n/N (%)

Table 4.2: Descriptive summary by sample time for variables: sex, event, treatment. Based on raw dataset. Each element is a sample for an individual.

Source: Own elaboration.

The total number of samples with a time block "out of time windows" is 101, which is a considerable amount. The design of the study also considers the scenario in which extra samples of the individuals are taken at discretionary timepoints, for instance, if a meerkat is to be euthanized an extra sample will be collected despise being out of predefined sample block. Only 4 out of these 101 "out of windows time blocks" are due to the former case, but samples out of predefined time blocks are allowed. We deem necessary to impute values to this variable for other future uses it. A self-programmed function `min_block_distance(x)` that receives a `Sample.Date` and returns the closest time Block to that sample date. For each observation with `Time.block = Out of time windows` the former function has been applied so as to impute data.

Another issue that arises is from `event` variable. Note that most the unclassified individuals are either in time Block 5 or out of time window block, thus it is risky to impute data to this variable as we might rise to a missclassification error which is a more advanced topic[9] we are not treating. All individuals with at least one unclassified state, will be removed from the analysis.

Note that at the beginning of the study (Block 1), the number of individuals in each treatment group is equal: 1 (33%), 2 (35%), 3 (33%)[10], and that those percentages evolve over time. This is what MSM will look at, all treatment groups contain the same number of individuals, what happens if one treatment succeeds but others fail? then the number of individuals in the coming time blocks will evolve depending on the success of the treatment, MSM will help us to identify if the dynamics is statistically significant or not.

Another issue coming from this data is the number of observations per individual. General MSM packages such as `msm` and `mstate`, require at least two observations per individual to be considered for the analysis, if an individual has just one observation this individual won't be considered for the analysis as we cannot infer anything from the dynamics of its life cycle since only one sample has been taken from it.

[9] If interested, one can refer to Hidden-Markov models for this issue.

[10] This is due to the random allocation of treatments defined in the design of the experiement.

Finally, after applying all the above changes, we get the cleaned dataset. The summary statistics for this dataset is shown below.

Descriptive statistics

Based on LifeSurvivalDataset after data cleaning process

	Block 1	Block 2	Block 3	Block 4	Block 5
Variable[1]	N = 64[1]	N = 112[1]	N = 164[1]	N = 87[1]	N = 46[1]
sex					
F	25 (39%)	38 (34%)	66 (40%)	39 (45%)	27 (59%)
M	39 (61%)	74 (66%)	98 (60%)	48 (55%)	19 (41%)
tx_grp					
1	19 (30%)	37 (33%)	38 (23%)	24 (28%)	10 (22%)
2	23 (36%)	34 (30%)	58 (35%)	30 (34%)	13 (28%)
3	22 (34%)	41 (37%)	68 (41%)	33 (38%)	23 (50%)
event					
1	43 (67%)	91 (81%)	81 (49%)	40 (46%)	15 (33%)
2	19 (30%)	11 (9.8%)	29 (18%)	19 (22%)	16 (35%)
4	2 (3.1%)	10 (8.9%)	49 (30%)	26 (30%)	3 (6.5%)
3			5 (3.0%)	2 (2.3%)	12 (26%)

[1] n/N (%)

Table 4.3: Descriptive summary by sample time for variables: sex, event, treatment. Based on cleaned dataset.

Source: Own elaboration.

After all the changes, the initial number of samples taken at Block 1 is reduced to 64 contrary to the raw dataset with 83 samples at the beginning of the study. One of the main reasons of individuals only having one sample taken throughout the study, is that some individuals that were sampled at time Block 1 were either euthanised or dead[11]. In the clean dataset, all individuals start in either state 1 or 2 (tuberculosis-free or tuberculosis infected), and transition to other different states throughout the coming time blocks, which is coherent. The number of individuals in each treatment group at time Block 1 now differs from the raw dataset: 1 (30%), 2 (36%), 3 (34%), though the number of samples in each treatment arm is not that big. We now present the sample size

[11] Remember that an individual is considered to have died if its not found for an interval time of 3 months since last visit.

Number of subjects at each time block *Based on dataset LifeSurvivalDataset*		
	Clean	Raw
Variable[1]	Subjects[1]	Subjects[1]
Time.block		
Block 1	60	79
Block 2	98	117
Block 3	134	163
Block 4	80	77
Block 5	43	74
Out of time windows		91
[1] Number of subjects at time block		

Table 4.4: Number of subjects for the cleaned and raw dataset

Source: Own elaboration.

First, note that the number of subjects at each time block evolves over time, increases and decreases depending on time block. This is nothing to worry about, number of individuals by time block might decrease due to entering an absorbing state. Similarly, if a meerkat is born within a treatment group at a given time block then it automatically enters the study at such time block.[12]. The ups and downs in the sample size is coherent with the design of the experiment.

We make use of `statetable.msm` function of `msm`, which will build a cross table with the initial state and the end state registered for each subject.

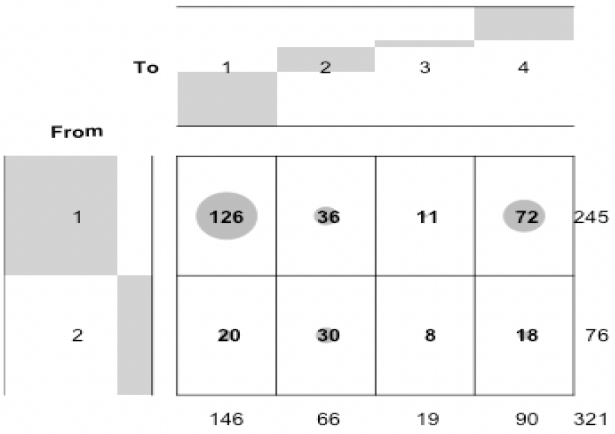


Figure 4.3: Number of transitions for cleaned dataset, from initial state to end state. (1)- Tuberculosis free; (2)- Tuberculosis infected; (3)- Euthanized; (4)- Dead due to other cause but euthanasia.

Source: Own elaboration.

Subjects either begin in state tuberculosis-free or tuberculosis-infected. For instance, from all 245 samples that began in state tuberculosis free, 72 of them resulted in dead due to other cause but euthanasia, 11 were euthanasied at the end of study period, 36 transitioned to tuberculosis-infected at the end of study, and 126 remained in the same state of tuberculosis-free. Note that, this plot only shows the initial state and the end state, but not the transitions within. This plot aims to identify which transitions should be allowed in the MSM diagram.

[12]The original design considers this scenario, in which case born wild meerkats are included in the study at the time block and treatment group they were born.

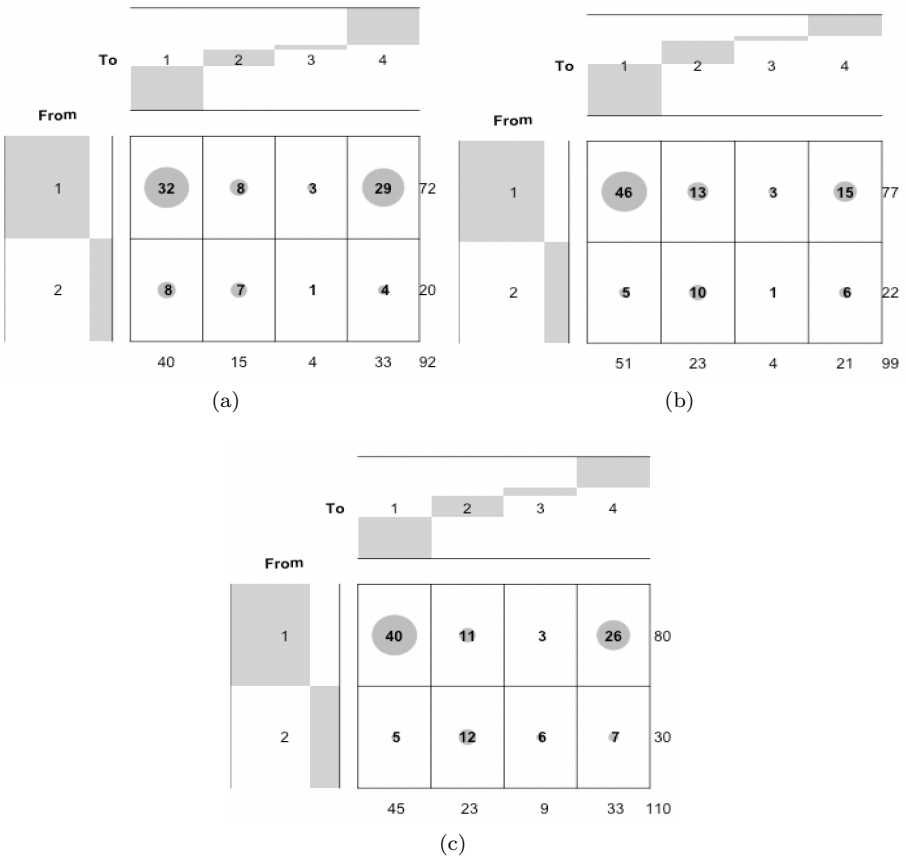


Figure 4.4: (a) Total number of transitions for treatment group 1 (high-susceptibility), (b) Total number of transitions for treatment group 2 (high-contact), (c) Total number of transitions for treatment group 3 (control).

At first glance, there seems to be a different between treatment group 1 and 2, the total number of samples in both treatment groups do not differ much (92 treatment group 1, 99 treatment group 2), although there seems to be a difference in the amount of transitions to death state: 33 in treament 1 group and 21 dead transitions in treatment 2, having the former more samples but less deaths. Treatment 3 group has the highest number of samples, and it's harder to assess the potential existence of a treatment difference, at first glance, with respect to other treatment groups. Once again, remember that this plot does not identify transitions between first and last samples, and thus we cannot guarantee this descriptive screnshoot holds throughout all the study.

4.2.5 Methodology

Model 1

The first MSM fitted will be composed of a state space $S = \{1, 2, 3, 4\}$. A bidirectional, time-homogeneous Markov, MSM is fitted. No censoring patterns are assumed, and two endpoints are defined: euthanasia (3), dead due to other cause but euthanasia (4). The following diagram depicts our MSM.

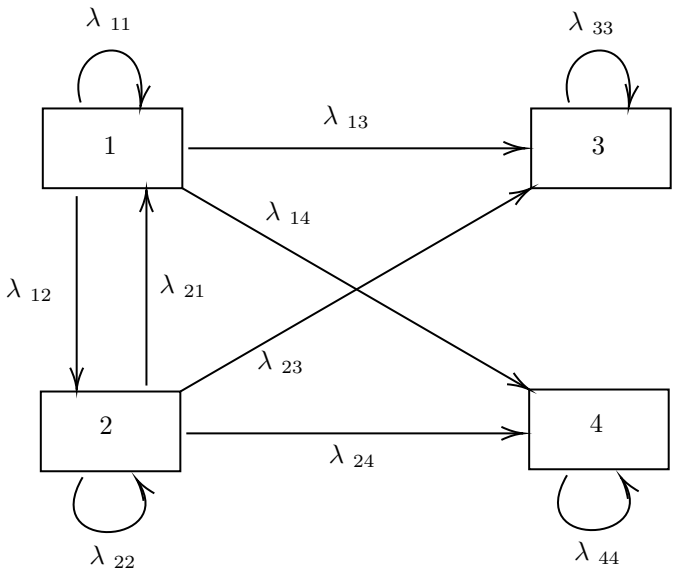


Figure 4.5: Multi-state diagram for the MSM proposed. (1)- Tuberculosis free; (2)- Tuberculosis infected; (3)- Euthanized; (4)- Dead due to other cause but euthanasia. Each λ_{kl} represent transition intensities for the transition $k \to l$.

Source: Own elaboration.

To fit the MSM model we will use `msm` package. The transition intensity matrix associated to figure 4.5 is:

$$Q_{|\mathbf{X}} = \begin{pmatrix} -\sum_{l\in\{2,3,4\}} \lambda_{1l} & \lambda_{12} & \lambda_{13} & \lambda_{14} \\ \lambda_{21} & -\sum_{l\in\{1,3,4\}} \lambda_{2l} & \lambda_{23} & \lambda_{24} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}_{|\mathbf{X}} \tag{4.1}$$

Two fixed-time covariates will be introduced in the regressions: sex and treatment group (tx_grp). A multiplicative functional form will be assumed for the regressions, hence a proportional hazards Cox model will be fitted. All the regressions will incorporate the former covariates, its expressions are of the form:

$$\lambda_{kl} = \lambda_{kl,0} \exp\{\beta_1 \text{sex} + \beta_2 \text{tx\_grp}\}$$

For each possible transition from state k to l. Although one can find `R` code attached in Appendix 4.2.6, an explanation of the most important parts of it will be detailed. The first step of the estimation process is to initialize the transition intensity matrix.

```
> Q <- rbind(c(0, 1/3, 1/3, 1/3),
+            c(1/3, 0, 1/3, 1/3),
+            c(0, 0, 0, 0),
+            c(0, 0, 0, 0)
+ )
```

The chosen initialization is arbitrary, one can initialize it to any values as long as values are coherent for a transition intensity matrix. Under `msm` package, diagonal entries are set to 0 (instead of minus the sum of the rows).

```
> Q.crude <- crudeinits.msm(event ~ start, id, data = data, qmatrix = Q)
```

Function `crudeinits.msm` gives a suitable initialized transition intensity matrix so as to guarantee convergence of the iterative method that will be used later. We have used variable `start` as time. Remember that under this MSM model, a global maximum is always guaranteed, no convergence analysis is needed based on initial condition. Now we get onto the main function to fit the MSM.

```
> model.msm <- msm(event ~ start, subject = id, exacttimes = TRUE,data = data,
+                  qmatrix = Q.crude, covariates = ~ tx_grp + sex,
+                  control = list(trace = 1, REPORT=1))
```

`event` $\sim$ `start` denotes that variable events is measured through time by variable *start*, subjects are denoted by variable *id*. Since we do not consider censoring, observed transitions are assumed to be known, that is why parameter `exacttimes` is set to `TRUE`. For all transitions we consider the same covariates, then we do not distinguish covariates by each regression hence `covariates = ~ tx_grp + sex`. Function `msm` runs `optim` function to maximise the minus log-likelihood of the MSM, to control the flux of the iterative algorithm and get information about state of convergence we set `control = list(trace = 1, REPORT=1)`, a listing for each iteration will be provided jointly with the minus log-likelihood at such. Finally, with the model set, convergence is always attained at iteration 77 with 39 gradient evaluations and a final value of the log-likelihood of 2491,96.

Model 2

In Model 1 we have splitted death state into two states, in this model we simplify the modelling by just considering three states $S = \{1, 2, 3\}$ by deleting the euthanasia state considered in Model 1. No censoring patterns are assumed, one endpoint is defined through state "dead". An illness-death Markov homogeneous MSM is fitted as next diagram shows.

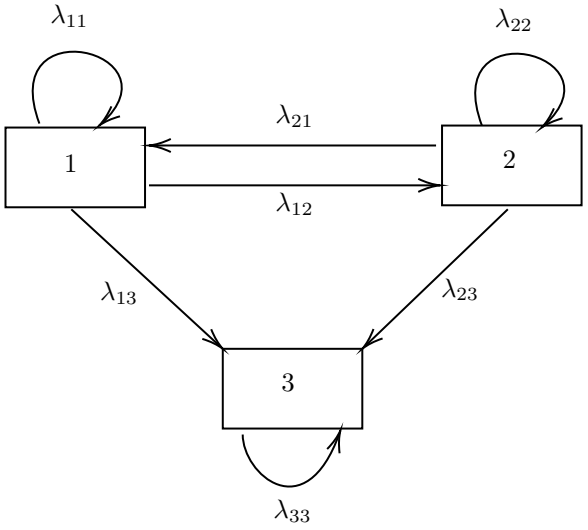


Figure 4.6: Multi-state diagram for the MSM proposed. (1)- Tuberculosis free; (2)- Tuberculosis infected; (3)- Dead. Each λ_{kl} represent transition intensities from $k \to l$.

Source: Own elaboration.

If we aim to test for treatment efficacy, by splitting death state as in Model 1 and given that the sample size is small we might decrease the power to test for treatment efficacy, for this reason we propose a simplification of Model 1. The transition intensity matrix for this model is shown.

$$Q_{|\mathbf{X}} = \begin{pmatrix} -\sum_{l\in\{2,3\}} \lambda_{1l} & \lambda_{12} & \lambda_{13} \\ \lambda_{21} & -\sum_{l\in\{1,3\}} \lambda_{2l} & \lambda_{23} \\ 0 & 0 & 0 \end{pmatrix}_{|\mathbf{X}} \tag{4.2}$$

The same covariates have been introduced as in Model 1, in all regressions fitted. A multiplicative functional form has been assumed so as to fit a proportional hazards Cox model. The code to fit this

Model 2 can be attached in the Appendix, the same reasoning for the code follows from explanation of Model 1. The convergence is attained at iteration 54 with 17 gradient evaluations and a final value of the log-likelihood of 2409,477.

4.2.6 Results

All relevant measures studied in the theoretical chapters: total sojourn time, mean sojourn time, transition probabilities, hazard ratios, etc, will be extracted from these MSM fitted to give answers to the questions posed.

Model 1

Reference category for variable sex has been set to female, whereas reference category for treatment variable has been set to the control group (`tx_grp` = 3). A baseline subject is such with `sex = F` and `tx_grp = 3` (a female from the control group).

Transition	**Parameter**	**EST (95% CI) [2]**			
		Baseline [1]	**Male**	**High-contact**	**High-susceptibility**
Tub. free → Tub. free	λ_{11}	-2.928e-03 (-3.540e-03, -2.422e-03)			
Tub. free → Tub. infected	λ_{12}	9.169e-04 (6.568e-04, 1.280e-03)	0.705 (0.3652, 1.361)	1.0659 (0.50577, 2.247)	6.756e-01 (2.833e-01, 1.611e+00)
Tub. free → Euthanized	λ_{13}	2.340e-04 (1.133e-04, 4.833e-04)	1.758 (0.4629, 6.673)	0.5098 (0.13087, 1.986)	1.740e-01 (2.141e-02, 1.415e+00)
Tub. free → Dead	λ_{14}	1.778e-03 (1.394e-03, 2.266e-03)	1.454 (0.8839, 2.392)	0.8855 (0.48634, 1.612)	1.357e+00 (7.912e-01, 6.006e+00)
Tub. infected → Tub.free	λ_{21}	1.769e-03 (1.120e-03, 2.794e-03)	1.367 (0.5518, 3.385)	1.0127 (0.32038, 3.201)	2.147e+00 (7.678e-01, 6.006e+00)
Tub. infected → Tub. infected	λ_{22}	-3.386e-03 (-1.262e-02, -9.086e-04)			
Tub. infected → Euthanized	λ_{23}	3.201e-05 (1.011e-63, 1.014e+54)	1.296 (0.3233, 5.198)	0.2036 (0.02494, 1.662)	3.931e-06 (1.099e-229, 1.406e+218)
Tub. infected → Dead	λ_{24}	1.585e-03 (9.751e-04, 2.576e-03)	1.005 (0.3955, 2.555)	1.6689 (0.60280, 4.620)	8.541e-01 (2.182e-01, 3.343e+00)

-2*log-likelihood : 2491.976

[1] Baseline is set for treatment as control group and sex as female
[2] Maximum likelihood estimates

Table 4.5: Maximum likelihood estimation of Model 1

Source: Own elaboration.

The MSM output shows different quantities at once; baseline hazards which are the hazards for a female meerkat that belongs to control group, and the hazard ratios for the levels of the covariates. All hazard ratios from every transition are shown with their respective confidence interval, those have been computed with a a 95% confidence but other confidence levels could have been chosen, next to the point estimate of the hazard. Notice that for all hazards and all transitions all confidence intervals contain the interval $[-1, 1]$, which means that with a 95% confidence we cannot assess neither a positive nor negative effect on the risks of transition given that the CI contain the values for which there's no effect. By looking at the output, no further conclusions can be made as no significant results are shown by the hazard ratios. Let's now look at the survival plots stratified by treatment group.

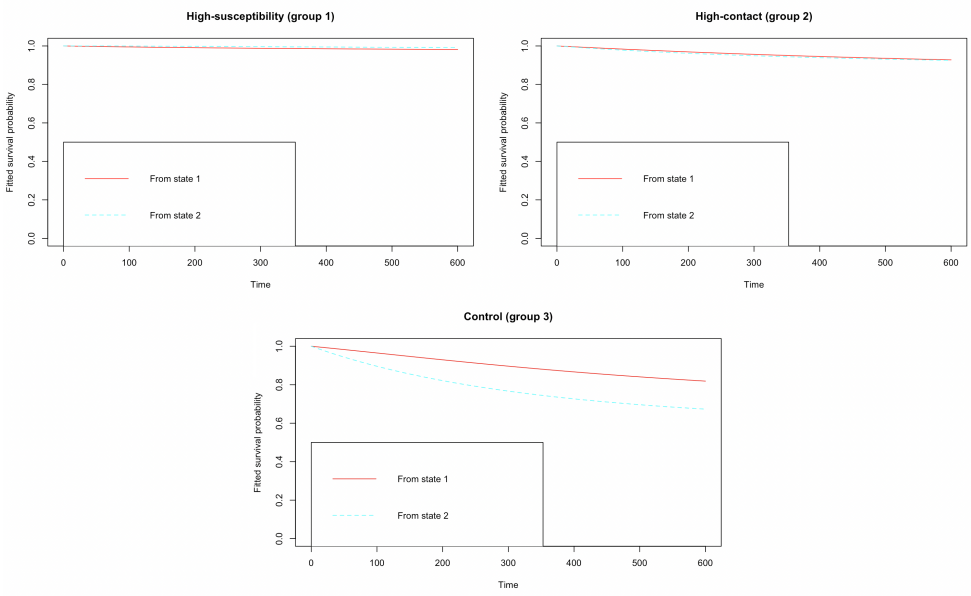


Figure 4.7: Survival plot from state (1) and (2) to euthanasia state (3).

Source: Own elaboration.

Survival plots represent the time until entering an absorbing state. Figure 4.7, shows the time until a meerkat is euthanasied based on previous state and stratifying by treatment group. It is clear that subjects belonging to high-susceptibility (HS) and high-contact (HC) have a higher survival time of being euthanasied than those in the control group. Further, note that the survival times for subjects in HS and HC groups are equal either if meerkat was previously infected by tuberculosis or tuberculosis-free. On the contrary, meerkats belonging to the control group have a less survivability with respect to being euthanasied, at day 600 more than 20% meerkats in the control group that were infected by tuberculosis were euthanasied. It is quite relevant that the survival curves for control group differ over time, meerkats infected by tuberculosis in the control group are more likely to be euthanasied (at every timepoint of the study) than those not infected by tuberculosis, this fact, though not conclusive, is indicative that the vaccination treatment might be effective as the control group experiences more euthanasia if infected by tuberculosis than other treatment groups. In fact, survival curves for HS and HC do not differ based on tuberculosis infection status. The main problem is we cannot conclude treatment is effective based on this argument, since we do not know the actual cause of the euthanasia: animals might have been selectively euthanasied and a biais might occur. Survival plots for the absorbing state ”dead due to other cause but euthanasia” is shown.

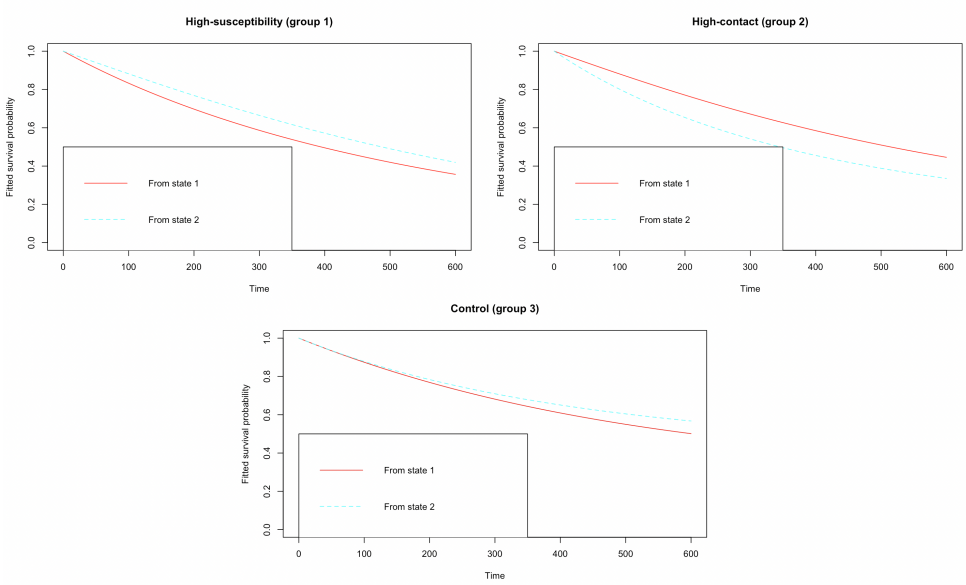


Figure 4.8: Survival plot from state (1) and (2) to dead state (3). Model 1.

Source: Own elaboration.

The survivability, with respect to the absobing state of death, between the treatment groups seem to differ from survival plot to euthanasia state. Now the control group seems to have a higher survivability: the time it takes to enter the death state is higher in the control group (without regard of initial state) than the HS and HC groups. This plots make it harder to extract a conclusion since it is contradictory to survival plots for euthanasia state. The decision to split the death state into two: euthanasia, dead due to other cause but euthanasia, make take a key role in this. The sample size is small, and by splitting the death state we might decrease the power to test for treatment differences, which is what might be causing this contradiction.

	State 1	State 2	State 3	State 4
High-susceptibility	459.13	101.02	Inf	Inf
High-contact	441.02	152.43	Inf	Inf
Control	393.33	126.18	Inf	Inf

Table 4.6: Total length of stay in each state by treatment group through all study period.

Source: Own elaboration.

In table 4.6 the total length of stay in each state, by treatment group, is presented. The total amount of time a meerkat from the HS treatment group is tuberculosis-free is 459 days, whereas HC treament group is 441 days and control group 393 days. Meerkats in the control group remain less days tuberculosis-free than meerkats in the other treatment groups. Furthermore, meerkats in the HS treatment group have the lowest number of days infected by tuberculosis (101 days), followed by the Control group (126 days), and the HC treatment group (152 days). Clearly, for absorbing states the total length of time is infinite by definition.

	State	Treatment	estimates	SE	L	U
1	State 1	High-susceptibility	364.39	77.52	240.15	552.91
3	State 1	High-contact	372.16	73.91	252.16	549.27
5	State 1	Control	341.66	68.36	230.82	505.73
2	State 2	High-susceptibility	266.39	102.49	125.32	566.25
4	State 2	High-contact	265.24	85.35	141.16	498.37
6	State 2	Control	262.43	69.91	155.68	442.37

Table 4.7: Mean sojourn time in each state by treatment group. State 1 = tuberculosis-free, State 2 = tuberculosis infected.

Source: Own elaboration.

The point estimate of the mean sojourn time, its standard error, and the lower and upper limits of the confidence interval at 95% confidence is shown in figure 4.7. The mean time a meerkat is expected to be free from tuberculosis in the HS treatment group is 364,39 days, whereas in the control group a meerkat is expected to be tuberculosis free for 341,66 days, though no further conclusions can be made as all confidence intervals overlap at 95% of confidence. Notice that there's a pattern on the variability of the groups, the HS treatment group always shows a higher variability, followed by HC group, and the control group, thus this might indicate that some heterogeneity patterns may arise between the groups, unfortunately we are not considering this in the modelling. Validation plots for the MSM fitted will be shown.

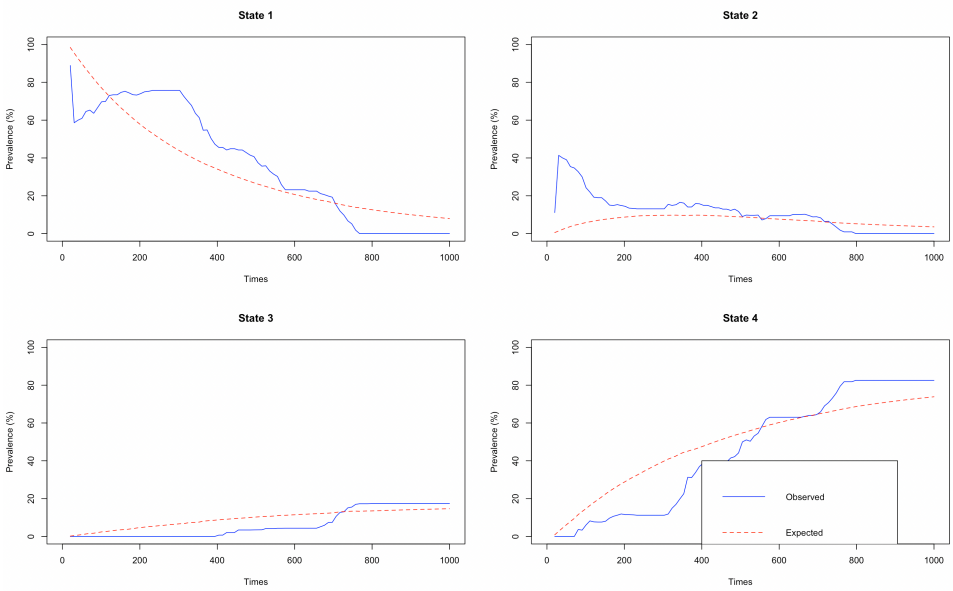


Figure 4.9: Validation plots for Model 1 for each state.

Source: Own elaboration.

This validation plots calculate the expected number of subjects in each state for the process and the observed ones, if the curves do fit then the model fits its theoretical counterpart, otherwise the model has deficiencies that should be tackled. Validation plot for state one shows the model has deficiencies estimating the number of subjects for this state, theres a long period of time [200, 700] at which our model overestimates the prevalence[13]. In death state we underestimate the number of subjects over the interval [0,600]. Overall, the model is not quite accurate based on the validation plots and some other actions should be taken to properly adjust the process. These discrepancies (theoretical vs empirical), could have many causes. One possibility is that the Markov model is non-homogeneous, or not even a Markov process. The censoring pattern assumed has been simplified and it is not correct (as it should be an interval-censored one), we have avoided introducing many time-varying covariates that could be key to explain the process, State 4 might be underestimated due to splitting the death state into State

[13]% of subjects in that state with respect to the total number of subjects in the process at that time.

3 and State 4, the model might account for heterogeneity[14], etc. Overall, with this model we cannot conclusively assess the efficacy of treatments.

[14]See frailty models for an extension of these.

Model 2

Reference category for variable sex has been set to female, whereas reference category for treatment variable has been set to the control group (`tx_grp` = 3). A baseline subject is such with `sex = F` and `tx_grp = 3` (a female from the control group).

Transition	Parameter	EST (95% CI) [2]			
		Baseline [1]	Male	High-contact	High-susceptibility
Tub. free → Tub. free	λ_{11}	-0.0029925 (-0.0036037, -0.002485)			
Tub. free → Tub. infected	λ_{12}	0.0009169 (0.0006568, 0.001280)	0.705 (0.3652, 1.361)	1.0659 (0.5058, 2.247)	0.6756 (0.2833, 1.611)
Tub. free → Dead	λ_{13}	0.0020756 (0.0016594, 0.002596)	1.489 (0.9345, 2.373)	0.8032 (0.4655, 1.386)	1.0988 (0.6647, 1.817)
Tub. infected → Tub.free	λ_{21}	0.0017689 (0.0011197, 0.002794)	1.367 (0.5518, 3.385)	1.0127 (0.3204, 3.201)	2.1474 (0.7678, 6.006)
Tub. infected → Tub. infected	λ_{22}	-0.0039526 (-0.0054038, -0.002891)			
Tub. infected → Dead	λ_{23}	0.0021837 (0.0014233, 0.003350)	1.088 (0.5008, 2.363)	0.9316 (0.4019, 2.160)	0.4206 (0.1198, 1.477)

-2*log-likelihood : 2406.593

[1] Baseline is set for treatment as control group and sex as female
[2] Maximum likelihood estimates

Table 4.8: Maximum likelihood estimation of Model 2.

Source: Own elaboration.

All hazards and all transition intensities all confidence intervals contain the interval $[-1, 1]$, which means that with a 95% confidence we cannot assess neither a positive nor a negative effect on the risk transitions given that the CI contains the values for which there's no effect. The only point estimate of the hazard ratios outside the $[-1, 1]$ interval is the effect of the HS treatment group from the transition tuberculosis infected to tuberculosis free, which could be interpreted as the recovery rate of tuberculosis once infected, but again the CI does not guarantee any statistical significant effect. By looking at the output, we cannot make further conclusions as no statistical significance can be derived from the estimates. We look for the survival plots now.

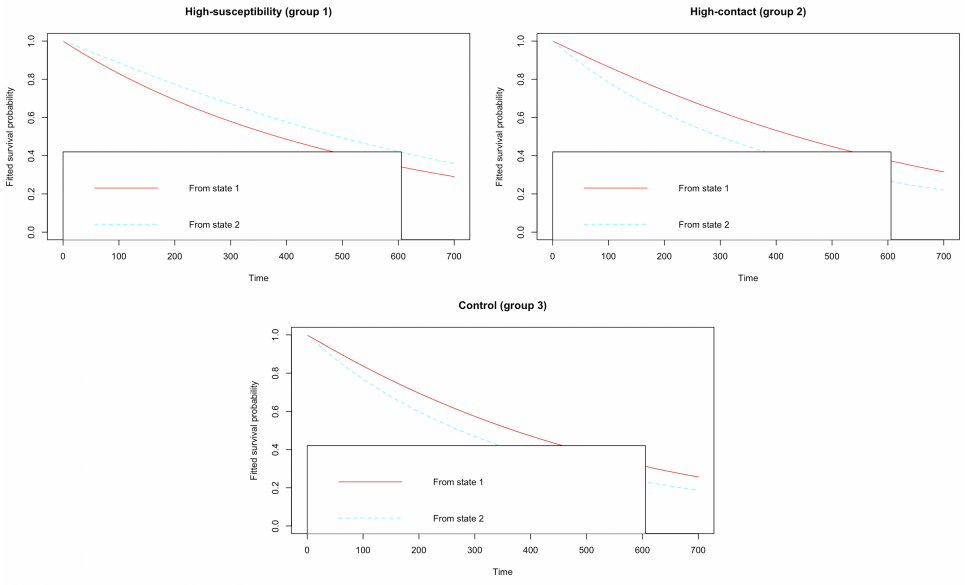


Figure 4.10: Survival plot from state (1) and (2) to euthanasia state (3). Model 2.

Source: Own elaboration.

Now we only have one absorbing state, then only one survival plot will be shown (Figure 4.10) from transience states (1) and (2) to the absorbing state death (3). The estimated survival for each treatment resembles the one of model 1. In the control group, meerkats infected by turberculosis have lower survivability than those tuberculosis-free, same goes for HC treatment group. For HS treatment group,

the argument is reversed which is contradictory, remember that this group has the highest variability among them all, and that we are not taking into account heterogeneity and this might be causing some contradictory results.

	State 1	State 2	State 3
High-susceptibility	466.03	104.12	Inf
High-contact	442.31	155.71	Inf
Control	389.37	123.03	Inf

Table 4.9: Total length of stay in each state by treatment group through all study period. Model 2.

Source: Own elaboration.

In table 4.9 the total length of stay in each state, by treatment group, is presented. The total length of state for the absorbing state death is infite by definition. The point estimates presented show that: HS is the treatment group where there's a highest amount of days without infection of tuberculosis and the less amount of days infected by tuberculosis, whereas the control group has the lowest number of days in a tuberculosis-free state and HC group the highest number of days with tuberculosis infection. Though, this results are non conclusive as we cannot determine statistical significance, those are only point estimates an no confidence intervals are presented yet. We will now present the mean sojourn times with their corresponding confidence intervals.

	State	Treatment	estimates	SE	L	U
1	State 1	High-susceptibility	367.53	77.44	243.19	555.45
3	State 1	High-contact	372.01	73.85	252.09	548.96
5	State 1	Control	339.33	67.23	230.14	500.33
2	State 2	High-susceptibility	270.51	103.52	127.77	572.70
4	State 2	High-contact	270.16	85.67	145.11	502.99
6	State 2	Control	258.48	67.22	155.26	430.33

Table 4.10: Mean sojourn time in each state by treatment group. State 1 = tuberculosis-free, State 2 = tuberculosis infected. Model 2.

Source: Own elaboration.

In table 4.10 the mean sojourn time in each state, by treatment group, is presented. All confidence intervals overlap at a 95% of confidence, thus no conclusions can be made out of this estimates, as the point estimates are not conclusive of any effect of the treatment in the length of stay in each state. Once again, note that the HS group has the highest variance of them all, whereas the control group has the less variability among all the three groups.

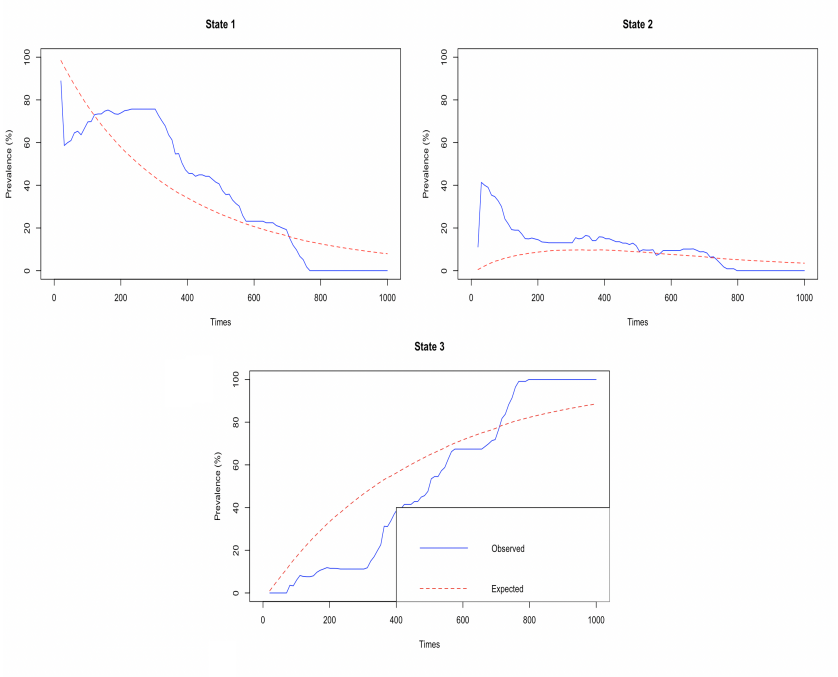


Figure 4.11: Validation plots for Model 2 for each state.

Source: Own elaboration.

Validation plots for model 2 resemble to those of model 1. The MSM fitted in both models have deficiencies, as the validation plots show that all states are either overestimated and/or underestimated throughout all the observed process. Then, we cannot assess our primary objectives of this analysis as results are not statistically significant. Furthermore, our results match with the original paper [14], as no statistically significant differences between HS treatment group and HC were found, but in our analysis we cannot even assess treatment differences between control and vaccination (which has been already found to be signficant in [14]).

Conclusions and further research

Data with multiple important survival endpoints are common, when such data arises standard survival analysis techniques such as Kaplan-Meier and Cox models turn obselete and Multi-state models become the best approach. Overall, one should consider a MSM approach when:

- There are multiple important survival endpoints that we must assess.
- Primary goal of the study involves understanding the dynamics of the process of interest.
- Some endpoints are recurrent. Which means that an individual can experience the same endpoint multiple times throughout the study period.
- There are competing events, which yields to a competing risk scenario.

MSM nurtures from complex longitudinal data with multiple endpoints, such data is common in other scientific disciplines rather than biostatistics, in acturial sciences and economics such data is frequent and so is in epidemiology. This thesis has focused on clinical biostatistical applications of such models, but those also find applications in the former scientific fields, and even in other diseases rather than tuberculosis such as Covid-19, HIV, and potentially emerging diseases like the smallpox monkey.

This thesis has successfully brought a clear and concise introduction to the theory of MSM driven by a homogeneous Markov process, which is the simplest stochastic process for the transitions one could consider, and yet theoretically enriching.

Details on the mathematical aspect of these models have been presented, and non-existent proofs, from the primary references, of important theorems and propositions have been provided such as: Proposition 1.3.2, Proposition 2.3.1, Proposition 1.1.1, Theorem 2.4.1, Theorem 2.4.2, Proposition 2.4.1. Furthermore, details and mathematical intuition on how to build a MSM likelihood function from scratch has also been presented. Other sections are fully on my own, due to an important lack of details in primary references, for instance section 2.5.

Our application was based on Patterson et al. [14], they aimed to study the dynamics of tuberculosis disease of Meerkats while considering multiple important endpoints (one of them was recurrent). Under such circumstances their statistical toolbox was mainly on Cox models. For this reason, we deemed necessary to propose a methodology based on MSM. No statistically significant results were found in the original paper regarding their primary goal: prove that trait-vaccination strategy based on social networks is more efficient than standard vaccination strategies. Despite the new statistical approach, we still could not conclusively assess statistically significant results. Note that, for the sake of simplicity, we made some crucial statistical assumptions: transitions driven by an homogeneous Markov process, all time-varying covariates were eliminated from the analysis, no censoring pattern. All these assumptions could have led to inconclusive results. As a further analysis, we could:

- Involve a MSM driven by other stochastic processses.
- Time-varying covariates.
- Interval censoring patterns.

Such extensions could enrich our analysis and potentially assess statistical significance.

Bibliography

[1] L.; de Leo G. Bolzoni L.; Real. "Transmission Heterogeneity and Control Strategies for Infectious Disease Emergence". In: *PLoS ONE* (2007). URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000747.

[2] T. Brezniak Z.; Zastawniak. "Basic stochastic processes". In: *London Barcelona: Springer* (2005).

[3] Christy Cassarly et al. "Assessing Type I error and power of multistate Markov models for panel data—A simulation study". In: *Communications in Statistics - Simulation and Computation* 46 (2017), pp. 7040 –7061.

[4] G.J.; Smith G.C.; Cheeseman C.L Delahay R.J.; Wilson. "Vaccinating badgers (Meles meles) against Mycobacterium bovis: The ecological considerations". In: *Vet J* (2003). URL: https://www.sciencedirect.com/science/article/pii/S1090023303000716?via%3Dihub.

[5] Julian A. Drewe. "Who infects whom? Social networks and tuberculosis transmission in wild meerkats". In: *Proceedings of the Royal Society B: Biological Sciences* 277.1681 (2010), pp. 633–642. DOI: 10.1098/rspb.2009.1775. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspb.2009.1775. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2009.1775.

[6] Ardo van den Hout. "Multi-state Survival Models for Interval-Censored Data". In: *Chapman and Hall/CRC* (2020).

[7] Christopher Jackson. "Multi-State Models for Panel Data: The msm Package for R". In: *Journal of Statistical Software* 38.8 (2011), 1–28. DOI: 10.18637/jss.v038.i08. URL: https://www.jstatsoft.org/index.php/jss/article/view/v038i08.

[8] Richard J.Cook and Jerald F.Lawless. "Multistate Models for the Analysis of Life History Data". In: (2018).

[9] G.F. Lawler. "Introduction to Stochastic Processes". In: *Boca Raton: Chapman Hall /CRC* (2006).

[10] Luís Meira-Machado et al. "Multi-state models for the analysis of time-to-event data". In: *Statistical Methods in Medical Research* 18 (2009), pp. 195 –222.

[11] Eugènia Negredo et al. "High risk and probability of progression to osteoporosis at 10 years in HIV-infected individuals: the role of PIs". In: *Journal of Antimicrobial Chemotherapy* 73.9 (June 2018), pp. 2452–2459. ISSN: 0305-7453. DOI: 10.1093/jac/dky201. eprint: https://academic.oup.com/jac/article-pdf/73/9/2452/25523204/dky201.pdf. URL: https://doi.org/10.1093/jac/dky201.

[12] J.R. Norris. "Markov Chains". In: *Cambridge : Cambridge University Press* (1997).

[13] Guadalupe Gómez; Klaus Langohr y Olga Julià. "Análisis de supervivencia". In: *Universitat Politècnica de Catalunya, Universitat de Barcelona* (2011).

[14] et.al Patterson SJ. "Trait-Based Vaccination of Individual Meerkats (Suricata suricatta) against Tuberculosis Provides Evidence to Support Targeted Disease Control." In: *Animals (Basel)* (2022). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8772857/.

[15] S.l. Resnick. "Adventures in stochastic processes". In: *Boston: Birkhauser* (1994).

[16] Antonin Slavík. "Product integration: its history and applications". In: (). eprint: https://www2.karlin.mff.cuni.cz/~prusv/ncmm/notes/download/product_integration.pdf.

[17] Isabelle Smith, Jane Nixon, and Linda Sharples. "Power and sample size for multistate model analysis of longitudinal discrete outcomes in disease prevention trials". In: *Statistics in Medicine* 40 (Feb. 2021). DOI: 10.1002/sim.8882.

[18] Rinku Sutradhar et al. "Multistate analysis of interval-censored longitudinal data: application to a cohort study on performance status among patients diagnosed with cancer." In: *American journal of epidemiology* 173 4 (2011), pp. 468–75.

[19] P. Todorovic. "An introduction to stochastic processes and their applications". In: *New York Barcelona: Springer* (1992).

Appendix

R code

Required packages and read data

```
library(msm)
library(dplyr)
library(mstate)
library(ggplot2)
library(gplots)
library(gt)
library(lubridate)
library(gtsummary)
webshot::install_phantomjs()

data <- read.csv("LifeTimeSurvival.csv", sep= ";", header = TRUE)
```

Data cleaning and descriptive statistics

```
# Generate states

# Function to create states
# State = 1, tuberculosis-free
# State = 2, infected of tuberculosis
# State = 3, dead.
#
    ##############################################################################


# If Death = NA or Positive = NA, we cannot determine state. State = 99
# If Death = 1 and/or Positive = 1/0/NA, then State = 3.
# If Death = 0 and Positive = 1, then State = 2.
# If Death = 0 and Positive = 0/NA, then State = 1.

classify <- function(death, positive, euth){
  if(euth == 1){
    return(3)
  }else{
    if((!is.na(death)) && (!is.na(positive))){
      if(death == 1){
        return(4)
      }else{
        if(positive == 1){
          return(2)
        }
        return(1)
      }
```

```
    }
  }
  return(99)
}

data$event <- mapply(classify, data$dth, data$Val.Pos, data$euth)

# GENERATE VARIABLE Time.Block
# Remember time blocks:
# Block 1: September-November 2014
# Block 2: December 2014-February 2015
# Block 3: July-October 2015
# Block 4: January-March 2016
# Block 5: July-September 2016

#Convert to datetime
data$Sample.Date <- gsub("/", "", data$Sample.Date)
data$Sample.Date <- dmy(data$Sample.Date)

timeblock <- function(x){
  date <- "Out of time windows"
  if(year(x) == 2014){
      if(9 <= month(x) && month(x) <= 11){
        date <- "Block 1"
      }
      if(month(x) == 12){
        date <- "Block 2"
      }
  }
  if(year(x) == 2015){
      if(1 <= month(x) && month(x) <= 2){
        date <- "Block 2"
      }
      if(7 <= month(x) && month(x) <= 10){
        date <- "Block 3"
      }
  }
  if(year(x) == 2016){
    if(1 <= month(x) && month(x) <= 3){
        date <- "Block 4"
    }
    if(7 <= month(x) && month(x) <= 9){
        date <- "Block 5"
    }
  }
  return(date)
}

data$Time.block <- sapply(data$Sample.Date, timeblock)
raw_data <- data
# Check which subjects are out of window due to euthanasia
table(data %>% filter(Time.block == "Out of time windows") %>% select(
    euth))

# Impute missing block times via minimum distance with respect to time
    windows

min_block_distance <- function(x){
```

```
  reference <- data.frame(initial = as.Date(c("2014-09-30", "2014-12-31"
    ,
                                              "2015-07-31", "2016-01-31"
                                                ,
                                              "2016-07-31")),
                          final = as.Date(c("2014-10-30", "2015-02-28",
                                            "2015-10-31", "2016-03-31",
                                            "2016-09-30")))
    min <- abs(as.Date(x)-reference[1,1])
    argfila_min <- 1
    for(i in 1:4){
      for(j in 1:2){
        dist <- abs(as.Date(x)-reference[i,j])
        if(dist < min){
          min <- dist
          argfila_min <- i
        }
      }
    }
  return(argfila_min)
}

rows <- as.numeric(rownames(data[data$Time.block == "Out of time windows
   ",]))
for(e in rows){
  block <- min_block_distance(data$Sample.Date[e])
  data$Time.block[e] <- paste("Block", as.character(block), sep = " ")
}


# MSM needs at least two observations per id to run, hence we will
   delete
# all subjects with only one observation scheme, as it does not
# offer any information

# Remove subjects with only one non-missing observed event

data <- data[data$event != 99,]
table_id <- table(data$id)

for(i in 1:length(table_id)){
  if(table_id[i] == 1){
    id_to_eliminate <- data[data$id == names(table_id[i]),]
    data <- data[!(data$id %in% c(id_to_eliminate)),]
  }
}

# Check for consistency as fixed time covariate, for each variable

check_fixed_time <- function(df, v){
  id <- unique(df$id)
  not_consistent <- c()
  for(e in id){
    aux <- df[df$id == e, v]
    t <- table(aux)
    if(nrow(t) >= 2){
      not_consistent <- c(not_consistent, e)
    }
```

```
  }
  return(not_consistent)
}

check_fixed_time(data, "sex")
check_fixed_time(data, "tx_grp")

# DESCRIPTIVE STATS AFTER AND BEFORE DATA CLEANING

summaries <- function(x, block){
  new_data <- x %>% subset(Time.block == block)
  new_data <- new_data[!duplicated(new_data, by = c(id, event)),]
  d<- new_data %>% select(sex, tx_grp, event) %>%
                   tbl_summary(statistic =
                                 c(sex, tx_grp, event) ~ "{n} ({p}%)")
  return(d)
}

d1_clean <- summaries(data, "Block 1")
d2_clean <- summaries(data, "Block 2")
d3_clean <- summaries(data, "Block 3")
d4_clean <- summaries(data, "Block 4")
d5_clean <- summaries(data, "Block 5")

merge_clean <-
  tbl_merge(
    tbls = list(d1_clean, d2_clean,d3_clean,d4_clean,d5_clean),
    tab_spanner = c("**Block 1**", "**Block 2**", "**Block 3**",
                    "**Block 4**", "**Block 5**")
  ) %>% modify_footnote(everything() ~ "n/N (%)") %>%
    modify_header(label = "**Variable**") %>%
    as_gt() %>%  tab_header(
    title = md("**Descriptive statistics**"),
    subtitle = md("*Based on LifeSurvivalDataset after data cleaning
       process*")
   )

gt::gtsave(merge_clean, file = file.path(getwd(), "clean_descriptive.png
   "))

d1_raw <- summaries(raw_data, "Block 1")
d2_raw <- summaries(raw_data, "Block 2")
d3_raw <- summaries(raw_data, "Block 3")
d4_raw <- summaries(raw_data, "Block 4")
d5_raw <- summaries(raw_data, "Block 5")
d6_raw <- summaries(raw_data, "Out of time windows")

merge_raw <-
  tbl_merge(
    tbls = list(d1_raw, d2_raw,d3_raw,d4_raw,d5_raw,d6_raw),
    tab_spanner = c("**Block 1**", "**Block 2**", "**Block 3**", "**
       Block 4**",
                    "**Block 5**", "**Out of time windows**")
  ) %>% modify_footnote(everything() ~ "n/N (%)") %>%
      modify_header(label = "**Variable**") %>%
      as_gt() %>% tab_header(
      title = md("**Descriptive statistics**"),
      subtitle = md("*Based on raw dataset LifeSurvivalDataset*")
```

```
    )

gt::gtsave(merge_raw, file = file.path(getwd(), "raw_descriptive.png"))

# Number of subjects at each time block

summaries_ind <- function(x){
  new_data <- x[!duplicated(x[c("id", "Time.block")]),]
  d<- new_data %>% select(Time.block) %>%
    tbl_summary(statistic =
                  c(Time.block) ~ "{n}")
  return(d)
}

subjects_clean <- summaries_ind(data)
subjects_raw <- summaries_ind(raw_data)

merge_clean_2 <-
  tbl_merge(
    tbls = list(subjects_clean, subjects_raw),
    tab_spanner = c("**Clean**", "**Raw**")
  ) %>% modify_footnote(everything() ~ "Number of subjects at time block
    ") %>%
  modify_header(label = "**Variable**", stat_0_1 = "Subjects", stat_0_2
    = "Subjects") %>%
  as_gt() %>% tab_header(
    title = md("**Number of subjects at each time block**"),
    subtitle = md("*Based on dataset LifeSurvivalDataset*")
  )

gt::gtsave(merge_clean_2, file = file.path(getwd(), "subjects.png"))

# Order by id and start time

data <- data[with(data, order(id, start)),]
raw_data <- raw_data[with(raw_data, order(id, start)),]

balloon_plot <- function(x, path, name){
  states.descriptive <- statetable.msm(event, id, data=x)
  dt <- as.table(as.matrix(states.descriptive))
  dev.copy(png,paste0(path, name))
        balloonplot(t(dt), main = "",xlab ="   To  ", ylab="From
              ", colsrt = 1,
              label = TRUE, show.margins = TRUE, dotcolor = "grey",
              scale.method = "volume", scale.range = "relative",
              colmar = 1, rowmar = 1.5)
  dev.off()
  return(1)
}

balloon_plot(data, getwd(), "/balloon_clean.jpg")

# Descriptive plots for treatment group
balloon_plot(data[data$tx_grp == 1,], getwd(), "/balloon_t1.png")
balloon_plot(data[data$tx_grp == 2,], getwd(), "/balloon_t2.png")
balloon_plot(data[data$tx_grp == 3,], getwd(), "/balloon_t3.png")
```

Fitting MSM

```
# FIT MODEL 1


# Which interval times correspond to each time block?

matrix <- matrix(data = NA, ncol = 2, nrow = 5)

for(i in 1:5){
  summary_start <- summary(data[data$Time.block ==
                                  paste("Block", as.character(i), sep =
                                      " "),
                                  "start"])
  summary_stop <- summary(data[data$Time.block ==
                                 paste("Block", as.character(i), sep = "
                                     "),
                               "stop"])
  matrix[i,1] <- summary_start[1]
  matrix[i,2] <- summary_stop[2]
}
matrix[1,1] <- 0

# Initialize transition intensity matrix
Q <- rbind(c(0, 1/3, 1/3, 1/3),
           c(1/3, 0, 1/3, 1/3),
           c(0, 0, 0, 0),
           c(0, 0, 0, 0)
)
Q.crude <- crudeinits.msm(event ~ start, id, data = data, qmatrix = Q)

# Fit model

# covariates as factors with reference level

data$sex <- relevel(factor(data$sex), ref = "F")
data$tx_grp <- relevel(factor(data$tx_grp), ref = 3)

data <- data[with(data, order(id, start)),]
model.msm <- msm(event ~ start, subject = id, exacttimes = TRUE,data =
    data,
                 qmatrix = Q.crude, covariates = ~ tx_grp + sex,
                 control = list(trace = 1, REPORT=1))

# MLE
model.msm

# Q matrix
q.matrix <- qmatrix.msm(model.msm)

# stratified q matrix by treatment

q.matrix.treatment <- list(qmatrix.msm(model.msm, covariates = list(tx_
    grp = 1)),
                           qmatrix.msm(model.msm, covariates = list(tx_
                               grp = 2)),
                           qmatrix.msm(model.msm, covariates = list(tx_
```

```
                                grp = 3))
                            )

# P matrix by time block
pmatrix.msm(model.msm, t = c(matrix[1,1], matrix[1,2]))
pmatrix.msm(model.msm, t = c(matrix[2,1], matrix[2,2]))
pmatrix.msm(model.msm, t = c(matrix[3,1], matrix[3,2]))
pmatrix.msm(model.msm, t = c(matrix[4,1], matrix[4,2]))
pmatrix.msm(model.msm, t = c(matrix[5,1], matrix[5,2]))

plot(model.msm)

# mean sojourn times

l1 <- cbind(data.frame(State = c("State 1", "State 2"),
                       data.frame(Treatment = "High-susceptibility"),
                       sojourn.msm(model.msm, covariates = list(tx_grp =
                           1))))
l2 <- cbind(State = c("State 1", "State 2"),
            data.frame(Treatment = "High-contact"),
            sojourn.msm(model.msm, covariates = list(tx_grp = 2)))
l3 <- cbind(State = c("State 1", "State 2"),
            data.frame(Treatment = "Control"),
            sojourn.msm(model.msm, covariates = list(tx_grp = 3)))
df <- rbind(l1,l2,l3, make.row.names = FALSE)
df <- df[order(df$State),]
xtable(df)

# total length of stay

l1 <- totlos.msm(model.msm, covariates = list(tx_grp = 1))
l2 <- totlos.msm(model.msm, covariates = list(tx_grp = 2))
l3 <- totlos.msm(model.msm, covariates = list(tx_grp = 3))

df <- rbind(l1,l2,l3)
rownames(df) <- c("High-susceptibility", "High-contact", "Control")
xtable(df)

# hazard ratios

hazard.msm(model.msm, cl = 0.95)

# survival plots

# euthanasia
plot(model.msm, from = c(1,2),to = 3, legend.pos = c(0,0.5),
     covariates = list(tx_grp = 1), range = c(0,600),
     main = "High-susceptibility (group 1)")
plot(model.msm, from = c(1,2), to = 3, legend.pos = c(0,0.5),
     covariates = list(tx_grp = 2), range = c(0,600),
     main = "High-contact (group 2)")
plot(model.msm, from = c(1,2), to = 3, legend.pos = c(0,0.5),
     covariates = list(tx_grp = 3), range = c(0,600),
     main = "Control (group 3)")

#death
plot(model.msm, from = c(1,2),to = 4, legend.pos = c(0,0.5),
     covariates = list(tx_grp = 1), range = c(0,600),
```

```
     main = "High-susceptibility (group 1)")
plot(model.msm, from = c(1,2), to = 4, legend.pos = c(0,0.5),
     covariates = list(tx_grp = 2), range = c(0,600),
     main = "High-contact (group 2)")
plot(model.msm, from = c(1,2), to = 4, legend.pos = c(0,0.5),
     covariates = list(tx_grp = 3), range = c(0,600),
     main = "Control (group 3)")

# model assessment

plot.prevalence.msm(model.msm, mintime= 0, maxtime= 1000)


# FIT MODEL 2

data$event <- ifelse(data$event == 4, 3, data$event)
Q <- rbind(c(0, 1/2, 1/2),
           c(1/2, 0, 1/2),
           c(0, 0, 0))
Q.crude <- crudeinits.msm(event ~ start, id, data = data, qmatrix = Q)

model.msm <- msm(event ~ start, subject = id,  data = data,
                 qmatrix = Q.crude, covariates = ~ tx_grp + sex,
                 control = list(trace = 1, REPORT=1), obstype = 2)

# MLE
model.msm

# Q matrix
q.matrix <- qmatrix.msm(model.msm)

# stratified q matrix by treatment

q.matrix.treatment <- list(qmatrix.msm(model.msm, covariates = list(tx_
   grp = 1)),
                           qmatrix.msm(model.msm, covariates = list(tx_
                              grp = 2)),
                           qmatrix.msm(model.msm, covariates = list(tx_
                              grp = 3))
)

# P matrix by time block
pmatrix.msm(model.msm, t = c(matrix[1,1], matrix[1,2]))
pmatrix.msm(model.msm, t = c(matrix[2,1], matrix[2,2]))
pmatrix.msm(model.msm, t = c(matrix[3,1], matrix[3,2]))
pmatrix.msm(model.msm, t = c(matrix[4,1], matrix[4,2]))
pmatrix.msm(model.msm, t = c(matrix[5,1], matrix[5,2]))

plot(model.msm)

# mean sojourn times

l1 <- cbind(data.frame(State = c("State 1", "State 2"),
                       data.frame(Treatment = "High-susceptibility"),
                       sojourn.msm(model.msm, covariates = list(tx_grp =
                           1))))
l2 <- cbind(State = c("State 1", "State 2"),
            data.frame(Treatment = "High-contact"),
```

```
            sojourn.msm(model.msm, covariates = list(tx_grp = 2)))
l3 <- cbind(State = c("State 1", "State 2"),
            data.frame(Treatment = "Control"),
            sojourn.msm(model.msm, covariates = list(tx_grp = 3)))
df <- rbind(l1,l2,l3, make.row.names = FALSE)
df <- df[order(df$State),]
xtable(df)

# total length of stay

l1 <- totlos.msm(model.msm, covariates = list(tx_grp = 1))
l2 <- totlos.msm(model.msm, covariates = list(tx_grp = 2))
l3 <- totlos.msm(model.msm, covariates = list(tx_grp = 3))

df <- rbind(l1,l2,l3)
rownames(df) <- c("High-susceptibility", "High-contact", "Control")
xtable(df)

# hazard ratios

hazard.msm(model.msm, cl = 0.95)

# survival plots

# to absorbing state
plot(model.msm, from = c(1,2),to = 3, legend.pos = c(0,0.42),
     covariates = list(tx_grp = 1), range = c(0,700),
     main = "High-susceptibility (group 1)")
plot(model.msm, from = c(1,2), to = 3, legend.pos = c(0,0.42),
     covariates = list(tx_grp = 2), range = c(0,700),
     main = "High-contact (group 2)")
plot(model.msm, from = c(1,2), to = 3, legend.pos = c(0,0.42),
     covariates = list(tx_grp = 3), range = c(0,700),
     main = "Control (group 3)")

# model assessment

plot.prevalence.msm(model.msm, mintime= 0, maxtime= 1000)
```