

Interuniversity Master in Statistics and Operations Research UPC-UB

Title: Interactive modelling and prognosis of a COVID-19 hospitalized patient via multistate models

Author: Leire Garmendia Bergés

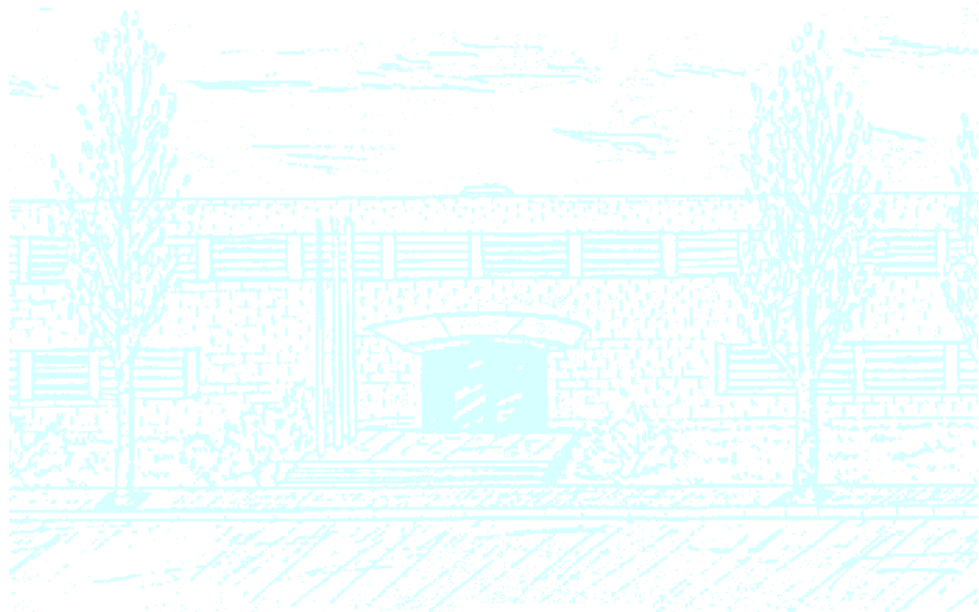
Advisor: Jordi Cortés Martínez

Co-advisor: Guadalupe Gómez Melis

Department: Statistics and Operational Research

University: Universitat Politècnica de Catalunya (UPC)

Academic year: 2021-2022



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Facultat de Matemàtiques i Estadística



Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Master Thesis

**Interactive modelling and prognosis of a
COVID-19 hospitalized patient via
multistate models**

Leire Garmendia Bergés

Advisor: Jordi Cortés Martínez

Co-advisor: Guadalupe Gómez Melis

Department of Statistics and Operational Research

Acknowledgements

Quisiera dar las gracias a Lupe y a Jordi por apoyarme en todo momento. Gracias Lupe por la capacidad que tienes de transmitirnos todo tu conocimiento. Gracias Jordi por siempre estar dispuesto a ayudarme y tranquilizarme cuando no veía una solución. También daros las gracias por darme la oportunidad de acercarme a la investigación tanto mediante el grupo GRBIO como mediante el proyecto DIVINE. Este tiempo con vosotros me ha ayudado a crecer tanto profesional como personalmente. Moltes gràcies!

Gracias Dani, Xavier, Klaus... por vuestro apoyo y por hacer los días difíciles más amenos.

Abstract

Keywords: Shiny app, multistate model, COVID-19

MSC2000: 200062N02 Survival analysis and censored data. Estimation.

MSC2000: 200092B15 Mathematical biology in general. General biostatistics.

MSC2000: 200062P10 Applications. Applications to biology and medical sciences.

MSC2000: 200068N99 Software. None of the above, but in this section.

Modelling the disease course regarding serious events and identifying prognostic factors is of great clinical relevance. Previous studies to predict high-risk critically ill cases among COVID-19 hospitalized patients have not yet arrived at a solid conclusion. Besides death, other intermediate events such as the need for invasive ventilation are relevant for clinical management. A team formed by clinicians and biostatisticians worked on the identification of the most clinically relevant states to explain the evolution of COVID-19 hospitalized patients, on the meaningful and plausible transitions between them, and on the characterization of the prognostic factors for those states. Based on this consensus, a multistate model (MSM) is proposed in order to learn about the disease progress. Motivated by this situation, an app is presented with two main goals: 1) to fit a MSM from specific data in a friendly way (programming skills are not required); 2) to predict the clinical evolution for a given patient based on the previous MSM. For the first objective, the user defines the states and transitions of the model as well as the covariates involved in each transition. The app returns descriptive information through histograms or barplots for the covariates, by box-plots to show the length of stay for each state and through instantaneous hazard plots to represent the risk of transition over time. For the second goal, information of the new patient at an initial state such as age or sex and the time for which predictions want to be made has to be provided. From these inputs, the app provides some indicators of the patient's evolution such as the probability of death or the most likely state at a fixed time. Furthermore, visual representations (e.g., the stacked transition probabilities plot) are given to make predictions more understandable. For illustrative purposes, we show how the app works using data from a multicohort study of more than 5,000 hospitalized adult COVID-19 patients from 8 Catalan hospitals during the first five waves of the pandemic. Different models have been fitted for the first Catalan pandemic wave, including as states the main outcomes –discharge and death– together with objective interventions during hospitalization such as non-invasive or invasive mechanical ventilation. The application and the underlying model are intended to be very useful for clinicians and to enhance the approach in modelling the course of other diseases with different stages of severity.

Notation

MSM	Multistate model
DIVINE	Dynamic evaluation of COVID-19 clinical states and their prognostic factors to improve the intra-hospital patient management
ICU	Intensive Care Unit
NIMV	Non-Invasive Mechanical Ventilation
IMV	Invasive Mechanical Ventilation
$k \rightarrow l$	Direct transition from state k to state l
HR	Hazard ratio
LS	Logarithmic score

Contents

Introduction	1
Multistate models	2
Software related to multistate models	3
Outline	5
Chapter 1. Multistate models	7
1. Introduction	7
2. Examples of multistate models	8
3. Steps of a multistate model	9
4. Characterization of a multistate model	10
5. Estimation of a multistate model	12
6. Prediction based on a multistate model	16
Chapter 2. MSMpred	19
1. Data	20
2. Model specification	20
3. Exploring the data	21
4. Fitted model	22
5. Graphics	24
6. Model validation	25
7. Predictions	25
8. Help	26
Chapter 3. Case study: DIVINE project	27
1. Data	29
2. Model specification	29
3. Exploring the data (EDA)	30
4. Fitted model	32
5. Graphics	35
6. Model validation	36
7. Predictions	38
Limitations and future work	41
References	45

Introduction

Modelling the disease course regarding serious events and identifying prognostic factors is of great clinical relevance. Since the first cases of coronavirus disease 2019 (COVID-19) in December 2019 in Wuhan (China), the entire world has dived into a pandemic that still continues. This pandemic presents a threat to global health with more than 6.2 million deaths due to COVID-19 [1]. Previous studies tried to predict high-risk critically ill cases among COVID-19 hospitalized patients, but they have not yet arrived at a solid conclusion. Besides death, other intermediate events such as the need for invasive ventilation are relevant for clinicians. The combination of all these events is required to learn about the progress of a COVID-19 patient.

Motivated by this situation, the *Dynamic evaluation of COVID-19 clinical states and their prognostic factors to improve the intra-hospital patient management* (DIVINE) project, funded by Generalitat de Catalunya (2020PANDE00148), aims to learn about the evolution of COVID-19 hospitalized patients. A team formed by clinicians and biostatisticians from Instituto de Investigación Biomédica de Bellvitge (IDIBELL) and Universitat Politècnica de Catalunya (UPC) works on this project with four main goals:

- (1) Identification of clinical relevant prognostic factors to severe pneumonia, need of mechanical ventilation, death or discharge in a cohort of hospitalized adult subjects with confirmed COVID-19.
- (2) Development and validation of a reliable clinical prediction tool for the early identification of potentially high-risk individuals among COVID-19 patients using multi-state model analysis.
- (3) Estimation of the COVID-19 incubation period in a cohort of hospitalized adult subjects.
- (4) Comparison of the clinical profile, the clinical management, and main outcomes of hospitalized adult subjects between the different waves.

For that, data from more than 5,000 hospitalized adult COVID-19 patients from 8 Catalan hospitals have been collected. This data is divided in four cohorts corresponding to four of the first five waves of the pandemic in Catalunya: March-April 2020 ($n_1 = 3460$), October-November 2020 ($n_2 = 516$), January-February 2021 ($n_3 = 637$) and July-August 2021 ($n_4 = 578$).

One of the objectives of the DIVINE project is to fit multistate models (MSM) to analyse the evolution of patients hospitalized due to COVID-19. Furthermore, we would like the clinicians to be able to use those models not only for that specific aim, but for any other disease or situation. Consequently, we understand there is a need for an intuitive and interactive tool that helps them to fit MSMs, but particularly to predict the evolution of a new individual basing on those models.

Motivated by this situation, in this Master's Thesis the second goal of that project is presented: the **MSMpred** shiny app. This app has two main goals: to fit a MSM from specific data and to predict the clinical evolution for a given patient based on the previous MSM. For that, only a subgroup of the first cohort of the DIVINE project is used, with data of the patients without ceiling of care (patients that do not have limitations on going to the different states).

Multistate models

During this COVID-19 pandemic several studies have been carried out to try to understand better this disease. Those studies have very diverse goals: analyse the mortality of patients with COVID-19, study the risk factors for going to intensive care unit (ICU), predict the need of ICU beds... Due to that, several statistical techniques were used depending on the purpose of each study.

One of the techniques that have been employed are MSMs. These models are very useful when the aim of the study is to describe the evolution of the individuals that have a disease with an increasing degree of severity, as in the case of COVID-19. One advantage of these models is that lot of aspects can be analysed, so these models not always are used in the same way. For example, let to see how the articles by Ursino et al. [2] and Mody et al. [3] used this methodology.

On the one hand, Ursino et al. [2] aimed to describe the evolution of patients admitted in the ICU due to COVID-19. For that, they consider different states depending on the type of mechanical ventilation, as well as three possible ways to get out of the ICU: ICU discharge, hospital discharge and death. Basing on that model, they analyse the clinical path of the patients since the admission on the ICU until 60 days after. Additionally, they looked for factors that could be related with transitions from one state to another, and analysed the effect of some drugs on the final outcome.

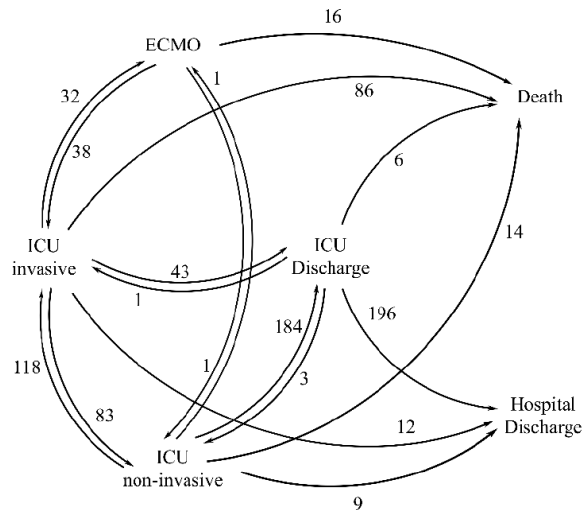


FIG. 1. MSM from Ursino et al.

On the other hand, Mody et al. [3] described the evolution of patients since the COVID-19 hospitalization, until they are discharged or die. The main objective

of this study was to characterize the clinical course of COVID-19 by means of estimating the length of stay into the hospital and in each state, by computing the cumulative incidences by 28 days since admission, and by analysing the association of some covariates and therapies on each transitions.

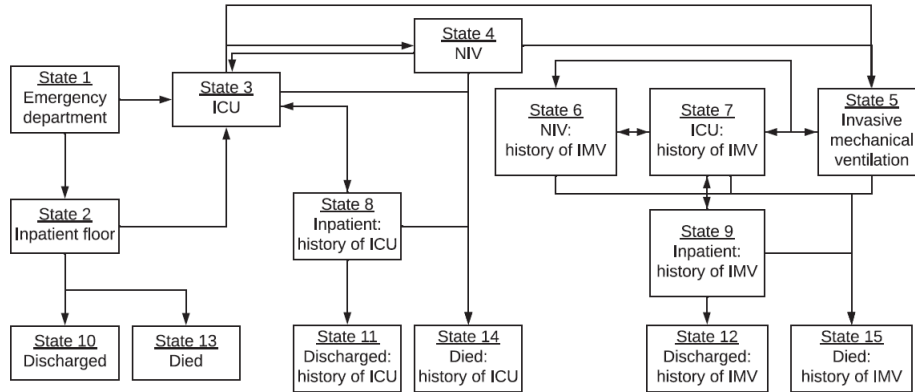


FIG. 2. MSM from Mody et al.

Those are only two examples that use MSMs to describe the evolution of COVID-19 patients. With those articles we have seen what have been done and which aspects we would like to analyse or improve. For example, we did not want neither to have bidirectional transitions as has the model in FIG. 1, nor to have several states representing the different absorbing states depending on the previous path of the individual as in FIG. 2.

MSMs are not only used to describe the evolution of individuals, they could also be used to make predictions on new individuals as have been made by Deschepper et al. [4]. They used multistate models to try to predict the number of needed beds in different sections of the hospital during the pandemic. With this aim, they divided the hospital wards in five states depending on whether there are COVID-19 patients in the ward or not and the cares needed by the COVID-19 patients. They construct a MSM with six states where the unique absorbing state was discharge (regardless of whether they go home or die).

Software related to multistate models

Before creating this app, we made a scoping review looking for different tools that allow clinicians to fit those models easily. We found several apps but, any of them has all the features that are important for us: fit a MSM and make predictions over new individuals. Due to that, we thought it would be interesting to create our own app taking the apps **MSM shiny** and **MSMplus shiny** as a reference.

We scrutinized those apps in order to make a list of aspects that we would like to include in our app or even improve them, as well as other characteristics that we would like to avoid. Before explaining our app, let us introduce you both apps and comment the positive and negative points that we found.

MSM shiny allows to fit different type of MSMs and it has some interesting outcomes statistically speaking.

Those are some positive characteristics that we would like to highlight:

- Possibility of adding an initial state that is not defined in the data.
- Define as many transitions as one wants.
- Assign different covariates to each transition.
- Fit different type of MSMs.
- Compare the fitted models.
- The presence of animation to show how individuals move between states.

But this app also has some aspects that we don't like or we want to improve:

- Once a transition is defined, it is not possible to remove it.
- The prediction part is quite poor.
- There is not a lot of information to help interpreting the results.
- The app does not allow to validate the model by means of a residual analysis.

MSMplus shiny is a more visual app, as it returns a variety of different graphs related with the fitted model.

Some of the positive aspects of this app are:

- Some theory of MSMs is explained and some indications about how to interpret the outputs are given.
- Possibility of changing the names of the states.
- The app is very friendly and it is plenty of graphical outputs.

Again, this app has some points to be improved:

- All the graphs are related with a few aspects of the model (e.g., transition probabilities, length of stay).
- There is only one covariate: age.
- It does not allow to fit different type of models.
- The app does not return a numeric summary of the fitted model.

Before starting to create our app, we made a list with some ideas that we would like to take into account in our app:

- Create a visual and friendly app easy to use.
- Possibility of defining all the possible transitions as well as to delete them.
- Fit different type of MSMs.
- Include the option to compare different models.
- Include a model validation.
- Predict the evolution of a new individual.
- Include a brief explanation of MSMs.
- Give some indications of how to interpret the results.

Regarding the packages available in R related with MSMs we found 23 packages, but the most important for us are `mstate` and `msm`.

Outline

This Master's Thesis is divided into an introduction, three main chapters, and the limitations and future work. The introduction has presented the reader to the project and its main objectives.

Chapter 1 provides the theoretical background used in the app. At the beginning the general ideas and examples for MSMs are presented, followed by the steps that need to be done before obtaining results from those models. Then, the different characterizations and estimations of the MSMs are explained. Finally, the reader can find some insights of the predictions based on MSMs.

Chapter 2 introduces in a global way the different sections that the user can find in **MSMpred**. Here the several inputs and outputs of our app are explained and some indications about how to interpret the results.

Chapter 3 presents a case study to illustrate the use of **MSMpred**. Data of the first cohort of the DIVINE project is used.

This work ends by listing the main limitations of **MSMpred** and some ideas that we would like to improve in the future.

Chapter 1

Multistate models

1. Introduction

A multistate model (MSM) is a model for a continuous time stochastic process allowing individuals to move among a finite number of states [9]. Within the scope of survival analysis, MSMs allow to describe complex clinical processes that change over time. Those models are formed by states and transitions, which represent, for instance, the different stages of a disease evolution and the possible paths to move between those states, respectively.

There are three different types of states: initial states are the ones where an individual could start the process; transient states are those in which individuals can get in and out of the state; and absorbing states are the ones where the process ends. Sometimes, states could also be both initial and transient, because people could start the process in these states but also people could arrive to them transitioning from other states.

As in classic survival analysis, the MSMs have some events of interest. In these models we focus on the transitions between states and the time until they occur. The most typical time-to-event analysis can be interpreted, as we will see, as a simple example of MSM, because when the event occurs, the individual makes the transition from one (initial) state to another state (of interest).

The main goals of MSMs are to:

- Understand the process of an individual or a group.
- Analyse the relationship between the covariates of interest and the process.
- Identify the risk factors for specific transitions.
- Develop predictive models for new individuals.

One relevant advantage of those models are that they allow to analyse the association between the individual characteristics and the propensity to make a transition in a specific time point (instantaneous hazard), being possible to relate different attributes of the individual to each transition.

There are three main steps to build a multistate model: 1) represent the clinical process by means of states and transitions; 2) decide which covariates or factors are considered in each transition; 3) fit the model. The first two steps usually require a clinical insight, so collaboration with medical experts could be helpful.

2. Examples of multistate models

Some examples, from the simplest to the most complex, are presented to better understand which type of clinical processes can be represented by a MSM and how to do it properly. For illustrative purposes, we are going to use some of the states considered in the DIVINE project, which we will describe in detail at the end of this section. In this project, we analyse the evolution of hospitalized patients due to COVID-19 in Catalan hospitals and the states describe several degrees of patient severity.

Example 1: progressive k-state model

The simplest example is the progressive 2-state model that is equivalent to the schema of the usual time-to-event analysis. We define two states, severe pneumonia (initial state) and death (absorbing state), with a single transition between them (FIG. 3). The aim of this model is to study the time to death for a hospitalized patient with severe pneumonia.

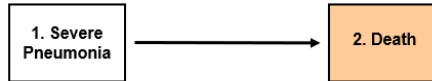


FIG. 3. Progressive 2-state model

It is possible to add more states to this model to form a progressive k -state model, $k \geq 2$. FIG. 4 shows a progressive 5-state model that represents one possible evolution that a patient hospitalized due to COVID-19 could have.

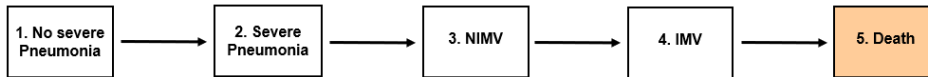


FIG. 4. Progressive 5-state model

In this progressive 5-state model, patients are hospitalized without severe pneumonia, and before dying (absorbing state), they go through the different transient states starting from no severe pneumonia, severe pneumonia, non-invasive mechanical ventilation (NIMV), and invasive mechanical ventilation (IMV). In this case, we have defined four transitions and, consequently, four events of interest.

Example 2: illness-death model

One common MSM is the illness-death model (FIG. 5), formed by three states (no severe pneumonia, severe pneumonia and death) and three transitions. The illness-death model is commonly used to study the incidence of a disease and to compare the rate of deaths between people with and without the disease. In this case, patients start the process without having severe pneumonia and they could go to severe pneumonia or die. If the patient dies (absorbing state) his process ends here, but if he gets severe pneumonia (transient state) his process continues.

Up to this point, all illustrations have had unidirectional transitions, but MSMs also allow to fit models with bidirectional transitions. One example is shown in FIG. 6. This model is very similar to the previous one, but in this case a patient that has severe pneumonia could recover to the no severe pneumonia state.

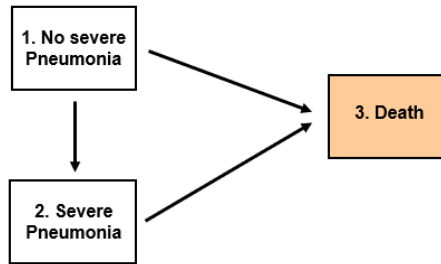


FIG. 5. Illness-death model

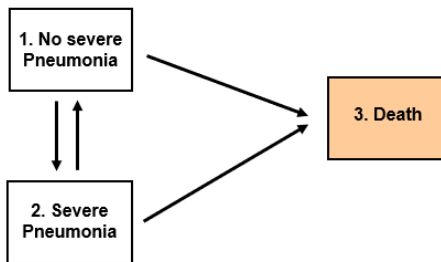


FIG. 6. Recursive illness-death model

Example 3: DIVINE model

The FIG. 7 shows the MSM considered in the DIVINE project. One of the goals of this project is to analyse the evolution of hospitalized patients admitted into hospital due to COVID-19. A team formed by clinicians and biostatisticians worked on the identification of the most clinically relevant states to explain the evolution of COVID-19 hospitalized patients, on the meaningful and plausible transitions between them, and on the characterization of the prognostic factors for those states. Based on this consensus, a multistate model with 7 different states (initial states: no severe pneumonia and severe pneumonia; transient states: NIMV, IMV and severe pneumonia recovery; absorbing states: discharge and death) and 14 transitions defined in the FIG. 7 is proposed in order to learn about the disease progress.

When a patient is admitted into the hospital, it can enter in one of the two initial states (no severe pneumonia or severe pneumonia). Then, during his/her hospitalization he/she makes different transitions and reaches some of the transient states (NIMV, IMV and severe pneumonia recovery) of the model. Finally, he/she is discharged or dies in hospital (absorbing states).

As you can observe the model can become increasingly complex as more states and transitions are added. The main challenge is to build the simplest model that has to be able to provide answers to the relevant clinical questions. In Chapter 5, this MSM will be presented as a case study.

3. Steps of a multistate model

As with any other model, different steps need to be done before obtaining results from the MSM (FIG. 8).

1. **Define the model:** define the states and transitions of the model.

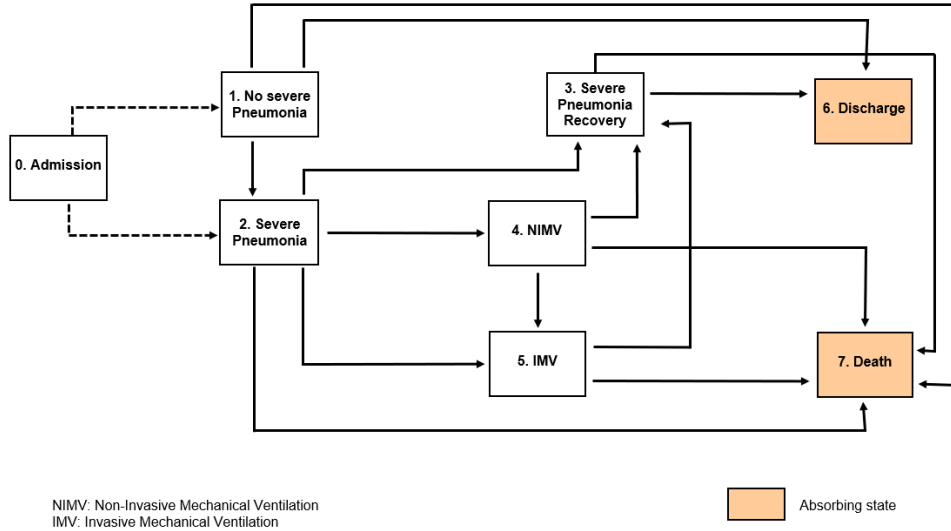


FIG. 7. DIVINE model

2. **Fit the non-parametric model:** fit a non-parametric model in order to make inference about the data.
3. **Include covariates:** following clinical criteria decide which covariates need to be included into the model. It is important to analyse the correlation between covariates, to avoid including related covariates.
4. **Covariate selection:** fit the full model and select the important covariates using backward stepwise selection.
5. **Model validation:** the assumptions on which the underlying model is based need to be checked. If those assumptions hold, the model is validated and it is possible to go to the last steps, otherwise, some changes need to be made: try to transform the numerical covariates, analyse interactions between covariates or include more covariates and fit the model including those changes, or redefine the model with the aim of obtaining a model that holds all the assumptions.
6. **Interpretation:** interpret the obtained results.
7. **Prediction:** predict the evolution of a new individual.

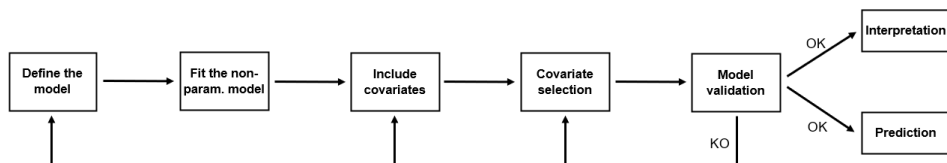


FIG. 8. Steps MSM

4. Characterization of a multistate model

There are different ways to characterize a MSM, related to each other in such a way that one characterization can be obtained from any of the others. The main characterizations of those models are based on transition probabilities, transition intensities or cumulative transition intensities. We have been inspired by the notation used by Cook and Lawless [10] to explain these characterizations.

The evolution of a patient could be understood as a continuous-time stochastic process formed by the different states that the patient visits over time. This stochastic process could be defined as $X = \{X(t) : t \geq 0\}$ where $X(t)$ represents the state in which the patient was at time t and takes values in the discrete set of states of the model, $\mathcal{R} = \{1, \dots, R\}$. Finally, the class $\mathcal{H}(t) = \{X(u), \mathbf{Z}(u), u \leq t\}$ contains the information of all the paths of all the individuals up to time t including the covariates $\mathbf{Z}(u)$ which can also be time-dependent.

- **Transition probability:** the probability of transition to state l at time t , provided that the patient was in state k at time s (for $s \leq t$) is defined as

$$\pi_{kl}(s, t; \mathcal{H}(s^-)) = \Pr\{X(t) = l | X(s) = k; \mathcal{H}(s^-)\}, \quad \forall k, l \in \mathcal{R} \quad (1)$$

- **Transition intensity:** the transitions intensities represent the probability of transition between two states, k and l , in a specific time point t and they are defined as

$$\begin{aligned} \lambda_{kl}(t; \mathcal{H}(t^-)) &= \lim_{\Delta t \rightarrow 0} \frac{\Pr\{X(t + \Delta t^-) = l | X(t^-) = k; \mathcal{H}(t^-)\}}{\Delta t} \\ &= \lim_{\Delta t \rightarrow 0} \frac{\pi_{kl}(t^-, t + \Delta t^-; \mathcal{H}(t^-))}{\Delta t}, \quad \forall k, l \in \mathcal{R} \end{aligned} \quad (2)$$

- **Cumulative transition intensity:** the cumulative transition intensity between states k and l is defined as

$$\Lambda_{kl}(t; \mathcal{H}(t^-)) = \int_0^t \lambda_{kl}(u; \mathcal{H}(u^-)) du, \quad \forall k, l \in \mathcal{R}. \quad (3)$$

The Markov property is met when

$$\Pr\{X(t) = l | X(s) = k; \mathcal{H}(s^-)\} = \Pr\{X(t) = l | X(s) = k\}$$

which implies that the future only depends on the present but not on the past.

Under the Markov assumption, the above expressions can be simplified:

- **Transition probability:**

$$\begin{aligned} \pi_{kl}(s, t) &= \Pr\{X(t) = l | X(s) = k\}, \quad \forall k \neq l \in \mathcal{R}, \\ \pi_{kk}(s, t) &= 1 - \sum_{k \neq l} \pi_{kl}(s, t), \quad \forall k \in \mathcal{R}. \end{aligned} \quad (4)$$

- **Transition intensity:**

$$\begin{aligned} \lambda_{kl}(t) &= \lim_{\Delta t \rightarrow 0} \frac{\Pr\{X(t + \Delta t^-) = l | X(t^-) = k\}}{\Delta t} \\ &= \lim_{\Delta t \rightarrow 0} \frac{\pi_{kl}(t, t + \Delta t^-)}{\Delta t}, \quad \forall k \neq l \in \mathcal{R}, \\ \lambda_{kk}(t) &= - \sum_{k \neq l} \lambda_{kl}(t), \quad \forall k \in \mathcal{R}. \end{aligned} \quad (5)$$

- **Cumulative transition intensity:**

$$\Lambda_{kl}(t) = \int_0^t \lambda_{kl}(u) du, \quad \forall k, l \in \mathcal{R}. \quad (6)$$

Those characterizations can be expressed in a matrix form using a $R \times R$ matrix for each (s, t) , (e.g., $\pi(s, t) = \{\pi_{kl}(s, t); k, l \in \mathcal{R} = \{1, \dots, r\}\}$).

Often we work under a Time-Homogeneous assumption. In this case, it is assumed that the transition intensities are constant along time, that is, $\lambda_{kl}(t) = \lambda_{kl}, \forall t$. This strong assumption allows to simplify the computation of the transition probabilities and work with a unique matrix $\pi = \pi(s, t), \forall s < t$.

Under the Time-Homogeneous Markov assumption, since $\lambda_{kl}(t) = \lambda_{kl}, \forall t$ and $\pi_{kl}(t, t + \Delta t) \approx \lambda_{kl}(t)\Delta t$, the transition probability between states k and l is:

$$\pi_{kl}(s, t) = \pi_{kl}(0, t - s) = \pi_{kl}(t - s), \quad s \leq u < t \quad (7)$$

which only depends on the elapsed time between leaving state k and reaching state l .

Using the Chapman-Kolmogorov equations the general computation of $\pi_{kl}(s, t), \forall s < t$ can be obtained using one-step probabilities $\pi_{kl}(u, u + 1)$ where $s \leq u < t$. The Chapman-Kolmogorov equations are:

$$\begin{aligned} \pi_{kl}(s, t) &= \Pr\{X(t) = l | X(s) = k\} = \frac{\Pr\{X(t) = l, X(s) = k\}}{\Pr\{X(s) = k\}} \\ &= \sum_{r=0}^R \Pr\{X(t) = l | X(u) = r\} \Pr\{X(u) = r | X(s) = k\} \\ &= \sum_{r=0}^R \pi_{rl}(u, t) \pi_{kr}(s, u), \quad s \leq u < t. \end{aligned} \quad (8)$$

Writing the Chapman-Kolmogorov equations in matrix notation:

$$\pi(s, t) = \pi(s, u) \pi(u, t), \quad s \leq u < t. \quad (9)$$

Consequently, under the Time-Homogeneous Markov assumption and because of the Chapman-Kolmogorov equations, a unique one-step transition probability matrix is needed.

The transition probability between states k and l can be interpreted as the probability of going through a specific transition $k \rightarrow l$ before time t , and the transition intensity between states k and l as the risk of going through a specific transition $k \rightarrow l$ in a specific time point t , respectively. However, the cumulative transition intensity is not possible to interpret.

5. Estimation of a multistate model

Once the MSM is defined, some non-parametric, semi-parametric and parametric estimations can be carried out. For those estimations, we will assume that the Markov assumption holds, and we will denote the direct transition from state k to state l as transition $k \rightarrow l^*$, being k and l contiguous states.

*Transitions between contiguous states.

5.1. Non-parametric estimation.

To estimate the cumulative transition intensities of the model, we can use the Nelson-Aalen estimator [10]. This estimator is based on the number of direct transitions $k \rightarrow l$ before time t , denoted by $N_{kl}(t)$, the number of individuals that go through the transition $k \rightarrow l$ at moment t , and the number of individuals in state k just before time t , denoted by $Y_k(t)$. Then, the Nelson-Aalen estimator for the cumulative transition intensity is

$$\begin{aligned}\hat{\Lambda}_{kl}(t) &= \int_0^t d\hat{\Lambda}_{kl}(u) = \int_0^t \frac{dN_{kl}(u)}{Y_k(u)}, \quad \forall l \neq k \in \mathcal{R} \\ \hat{\Lambda}_{kk}(t) &= - \int_0^t \hat{\Lambda}_{kl}(t), \quad \forall k \in \mathcal{R},\end{aligned}\tag{10}$$

where $d\hat{\Lambda}_{kl}(t) = \hat{\lambda}_{kl}(t)du$ and $Y_k(t) = \sum_{i=1}^n Y_{i,k}(t)$.

To estimate the transition probabilities of the model, we can use the Aalen-Johansen estimator [10]. This estimator is based on the previously estimated cumulative transition intensities denoted as $d\hat{\Lambda}(u)$. Then, the Aalen-Johansen estimator for the transition probability is

$$\hat{\pi}(s, t) = \prod_{s < u \leq t} \{I + d\hat{\Lambda}(u)\}\tag{11}$$

where $d\hat{\Lambda}(u)$ is a $R \times R$ matrix with elements $d\hat{\Lambda}_{kl}(u)$, $\forall k, l \in \mathcal{R}$, estimated via Nelson-Aalen estimator.

5.2. Semi-parametric estimation.

The Cox or proportional hazards model provides the most usual semi-parametrical estimations of the MSM. They are semi-parametric because the baseline transition intensities are not following a specific parametric distribution. For those models, transition intensities are defined as in equation (5). One of the main benefits of the Cox models is that they allowed to relate the characteristics of an individual (described by some covariates) and the transition intensities:

$$\lambda_{kl}(t; \mathbf{Z}) = \lambda_{kl,0}(t) \exp(\beta_{kl}^T \mathbf{Z})\tag{12}$$

where $\lambda_{kl,0}(t)$ is the baseline intensity function for the transition $k \rightarrow l$, β_{kl} is the vector of regression parameters, and $\mathbf{Z} = (Z_1, \dots, Z_p)$ is the covariate vector.

The association of a covariate Z_q with the transition intensity of a specific transition $k \rightarrow l$ can be measured by means of hazard ratios (HR):

$$\text{HR}_{kl,q} = \frac{\lambda_{kl}(t; \mathbf{Z} = (0, \dots, 0, Z_q = 1, 0, \dots, 0))}{\lambda_{kl}(t; \mathbf{Z} = (0, \dots, 0, Z_q = 0, 0, \dots, 0))} = \exp(\beta_{kl,q}).\tag{13}$$

We say that the covariate Z_q is a risk factor for this specific transition if $\text{HR}_{kl} = \exp(\beta_{kl,q}) > 1$ and a protective factor if $\text{HR}_{kl} = \exp(\beta_{kl,q}) < 1$ or what is the same, risk factor if $\beta_{kl,q} > 0$ and protective factor if $\beta_{kl,q} < 0$. If $\text{HR}_{kl} = \exp(\beta_{kl,q}) = 1$, we say that the covariate Z_q is not associated with the intensity of the transition $k \rightarrow t$.

After defining the model, it is necessary to estimate the parameters β_{kl} and the baseline hazard function $\lambda_{kl,0}(t)$. For that, the partial likelihood, $L(\beta)$, and the likelihood function conditioned to the estimated values $\hat{\beta}_{kl}$ need to be maximized.

To obtain the Cox partial likelihood $L(\boldsymbol{\beta})$ we need some information. For each individual we have the follow up time, τ_i , the multistate process, $X_i(t)$, and the transition or censoring times for each transition, $t_{i,kl}$. For each transition, we have the risk set, $R_k(t_{i,kl})$, with the individuals at risk in state k at time $t_{i,kl}$ and the covariate vector of the individuals at risk, \mathbf{Z}_j . Then the Cox partial likelihood is defined as

$$L(\boldsymbol{\beta}) = \prod_{k \rightarrow l} \prod_{i=1}^n \frac{\exp(\boldsymbol{\beta}_{kl}^T \mathbf{Z}_i)}{\sum_{j \in R_k(t_{i,kl})} \exp(\boldsymbol{\beta}_{kl}^T \mathbf{Z}_j)}. \quad (14)$$

The estimations of the coefficients, $\hat{\boldsymbol{\beta}}_{kl}$, can be obtained maximizing the Cox partial likelihood (14).

After that, the estimations $\hat{\boldsymbol{\beta}}_{kl}$ are used to estimate the baseline hazard function, $\lambda_{kl,0}(t)$. There are several ways to perform this estimation, but we will use the Breslow's estimate. Assuming that there are r failure times for a specific transition $k \rightarrow l$, $t_{kl}^{(1)} < \dots < t_{kl}^{(r)}$, the number of failures at each $t_{kl}^{(j)}$ are d_{kl}^j and the number of individuals at risk n_{kl}^j , the Breslow's estimator for the baseline hazard function at time $t_{kl}^{(j)}$ is

$$\hat{\lambda}_{kl,0}^B(t_{kl}^{(j)}) = \frac{d_{kl}^j}{\sum_{h \in R_k(t_{kl}^{(j)})} \exp(\hat{\boldsymbol{\beta}}_{kl}^T \mathbf{Z}_h)}. \quad (15)$$

Following the same idea, is possible to estimate the Breslow's estimator for the cumulative baseline hazard function for $t_{kl}^{(m)} \leq t_{kl} \leq t_{kl}^{(m+1)}$, $m = 1, \dots, r-1$:

$$\hat{\Lambda}_{kl,0}^B(t_{kl}) = \sum_{j=1}^m \frac{d_{kl}^j}{\sum_{h \in R_k(t_{kl}^{(j)})} \exp(\hat{\boldsymbol{\beta}}_{kl}^T \mathbf{Z}_h)}. \quad (16)$$

After fitting the Cox model, some assumptions should be graphically validated using several types of residuals:

- **Martingale-based residuals.** To validate linearity in continuous variables, we explore the behaviour of the residuals of the model fitted without the assessed covariate against the same covariate.
- **Residuals based on the scores.** To validate the global fit and to detect influential individuals.
- **Schoenfeld residuals.** To validate the proportional hazards premise.

Those assumptions need to be checked for each transition of the model. So, the different type of residuals need to be computed for each transition. Now, we are going to formally define those residuals.

The **martingale-based residuals** are defined for each individual at each transition. They represent the difference between the number of observed events and the number of expected events from the Cox model. For each individual $i = 1, \dots, n$, the martingale-based residuals associated to transition $k \rightarrow l$ are

$$r_{M_{i,kl}} = \begin{cases} 1 - \exp(\hat{\boldsymbol{\beta}}_{kl}^T \mathbf{Z}_i) \hat{\Lambda}_{kl,0}^B(y_i), & \text{if } \delta_{i,kl} = 1 \\ 0 - \exp(\hat{\boldsymbol{\beta}}_{kl}^T \mathbf{Z}_i) \hat{\Lambda}_{kl,0}^B(y_i), & \text{if } \delta_{i,kl} = 0. \end{cases} \quad (17)$$

where $\delta_{i,kl}$ represents the transition indicator that take value 1 if the individual i makes the transition $k \rightarrow l$, and 0 otherwise. They take values between $-\infty$ and

1 in the case of non-censored observations and between $-\infty$ and 0 in the case of censored observations. Therefore, they are not symmetrically distributed around 0.

These residuals serve to determine the best transformation for a covariate in such a way that it optimally explains the time to an individual passes through a certain transition. To find the best transformation for the covariate Z_q in the transition $k \rightarrow l$, the martingale-based residual from a Cox model adjusted with the other $p - 1$ covariates need to be computed. Then, the graphic of residuals $r_{M_{i,kl}}$ respect to the value of the covariate $Z_{i,q}$ are represented with a smoothed curve of the points trajectory along the x-axis. If the smoothed curve is reasonably linear, the covariate Z_q does not require any transformation in the transition $k \rightarrow l$.

Residuals based on the score are defined on a particular transition $k \rightarrow l$ for each covariate Z_q and for each individual i . For each $i = 1, \dots, n$ and $q = 1, \dots, p$, the residuals based on the score are

$$r_{S_{i,kl,q}}(t) = \int_0^t \{Z_{i,q}(s) - \bar{Z}_q(s)\} d\hat{M}_{i,kl}(s) \quad (18)$$

where

$$\begin{aligned} \bar{Z}_q(t) &= \frac{\sum_{i=1}^n J_{i,kl}(t) Z_{i,q} \exp\{\boldsymbol{\beta}_{kl}^T \mathbf{Z}_i(t)\}}{\sum_{i=1}^n J_{i,kl}(t) \exp\{\boldsymbol{\beta}_{kl}^T \mathbf{Z}_i(t)\}} \\ \hat{M}_{i,kl}(t) &= N_{i,kl}(t) - \int_0^t J_{i,kl}(s) \exp\{\boldsymbol{\beta}_{kl}^T \mathbf{Z}_i(s)\} d\hat{\Lambda}_0^B(s) \end{aligned} \quad (19)$$

being $J_{i,kl}(t) = \mathbf{1}\{\text{individual } i \text{ is at risk for transition } k \rightarrow l \text{ before } t\}$ the risk indicator and $N_i(t)$ the transition indicator for each individual i .

We plot those residuals based on the score versus $Z_{i,q}$ to determine the influence of the individual i in the estimation of the coefficients of the transition $k \rightarrow l$. That is, those residuals represent the difference between the estimator obtained when adjusting the Cox model for the transition $k \rightarrow l$ considering all the individuals, $\hat{\boldsymbol{\beta}}_{kl}$, and the estimator from the model without taking into account the individual i , $\hat{\boldsymbol{\beta}}_{kl(i)}$. So, those individuals far away from the others have a higher influence on the model estimates.

In practice instead of using the residuals based on the score some transformations like the dfbeta and dfbetas residuals are used. The dfbeta residuals also give a measure of the approximate change of the coefficients if the individual i is not taken into account [11]:

$$r_{df_{i,kl}}(t) = \hat{\boldsymbol{\beta}}_{kl} - \hat{\boldsymbol{\beta}}_{kl(i)} \quad (20)$$

The dfbetas residuals are the standardized dfbeta residuals.

Finally, the **Schoenfeld residuals** for an explicit transition $k \rightarrow l$ are defined for each covariate q and for each individual i . For each $i = 1, \dots, n$ and $q = 1, \dots, p$, they are defined as

$$r_{SC_{i,kl,q}} = \delta_{i,kl} J_{i,kl}(t) \{Z_{i,q} - \bar{Z}_q(T_i)\} \quad (21)$$

The Schoenfeld residuals determine the difference between the observed and expected value of the covariate Z_q in each transitioning time from state k to state l .

The graphic of residuals $r_{SC_{kl,q}}$ for each individual are represented with a smoothed curve of the points and the line $r_{SC_{kl,q}} = 0$. If the confidence interval of the

smoothed curve covers the line $r_{SC_{kl,q}} = 0$, the proportionality of the hazard is not severely violated.

Despite the residuals based on the score and the Schoenfeld residuals have been defined taking into account time-dependent covariates, those definitions can be simplified when only basal covariates are taken into account.

5.3. Parametric estimation.

We can also assume parametric distributions in the context of MSM [12]. Unlike the non- or semi-parametric estimation, in the parametric MSM, the baseline hazard function for the transition $k \rightarrow l$, $\lambda_{kl,0}(t)$, follows a parametric distribution such that $\lambda_{kl,0}(t) > 0$. We can use several distributions such as exponential, Weibull, Gompertz, log-logistic, log-normal, among others.

The main drawback of this type of estimation is that the assumption that the data follows a precise distribution need to be done. Moreover, if we want to consider different distributions for each transitions, we need to model those transitions separately.

In our present work we disregard this option.

6. Prediction based on a multistate model

As with any type of model, MSM may be used to make predictions for a new individual. For that, some characteristics of the patient such as the observed history of states and covariates until time t_0 and the initial state, $\mathcal{H}(t_0) = \{\mathcal{X}(t_0), \mathcal{Z}(t_0)\}$, need to be known. Based on that and in a previously fitted model, the future process, $\{X(s); s > t_0\}$, could be predicted.

In order to obtain those predictions for time $t_1 > t_0$, a predictive model needs to be specified to obtain $\tilde{P}\{X(t_1) = x|\mathcal{H}(t_0)\}$. The tilde indicates that this is the predicted probability and not the observed probability, $\Pr\{X(t_1) = x|\mathcal{H}(t_0)\}$. Once the predictive model is obtained, the transition probabilities, hazards functions... could be obtained in order to forecast different aspects of interest. For example, the probability of being in each state after time t could be deducted from the transition probabilities for the new patient.

To assess how good are those predictions, we need to analyse the calibration and sharpness of the predictive model. Concerning the calibration analysis the predicted probabilities, \tilde{P} , and the true probabilities, P , are compared, aiming to check how near/far are from each other. Ultimately systematically biased predictions can be detected. With the sharpness of the model we analyse if initial conditions or covariates, $\mathcal{H}(t_0)$, are highly predictive of the state an individuals will be in later.

There are some scoring rules that combine both aspects, calibration and sharpness, to analyse the performance of the predictive models like the logarithmic score also known as Kullback-Leibler score.

The logarithmic score for a given individual i is computed as

$$\text{LS}_i(\tilde{P}, t_1) = -\log \tilde{P}\{X_i(t_1) = x_i(t_1)|\mathcal{H}_i(t_0)\} \quad (22)$$

where $x_i(t_1)$ is the observed state at time t_1 and individual i is not included in the dataset used to fit the model.

If the predicted performance of the model wants to be analysed for a group of n individuals, the following formula can be used:

$$\begin{aligned} \text{LS}(\tilde{P}, t_1) &= -\frac{1}{n} \sum_{i=1}^n \text{LS}_i(\tilde{P}, t_1) \\ &= -\frac{1}{n} \sum_{i=1}^n \log \tilde{P}\{X_i(t_1) = x_i(t_1) | \mathcal{H}_i(t_0)\} \end{aligned} \quad (23)$$

We divide by n because otherwise the score always increases when more individuals are analysed. The logarithmic score takes values between 0 and ∞ .

That score is not interpretable, but it is useful when some models want to be compared. Let's assume that we have two models and that for the individual i model 1 has $\tilde{P}_1\{X_i(t_1) = x_i(t_1) | \mathcal{H}_i(t_0)\} = p_1$ and model 2 $\tilde{P}_2\{X_i(t_1) = x_i(t_1) | \mathcal{H}_i(t_0)\} = p_2$, being $p_1 < p_2$. As the probability of guessing the observed state is measured, the model with a higher probability is chosen, that is, model 2. But, when computing the logarithmic score we obtain the values $\text{LS}(\tilde{P}_1, t_1) = -\log p_1$ and $\text{LS}(\tilde{P}_2, t_1) = -\log p_2$, not p_1 and p_2 . As the logarithmic function is an increasing function, $\log p_1 < \log p_2$ so we prefer the model with a higher value, but when the logarithm is multiplied by -1 , $-\log p_1 > -\log p_2$, so lower values of $\text{LS}(\tilde{P}, t_1)$ are preferable.

If instead of just comparing the predictive performance of two models for one individual, we want to make the comparison for a group of individuals, once more the model with a lower logarithmic score needs to be chosen. This is because when several individuals are analysed, we sum the $\text{LS}_i(\tilde{P}, t_1)$ of each individual and divide it by the total number of individuals analysed, and both functions are increasing functions. Consequently, when comparing different models if the one with the best predictive performance wants to be selected, the model with a lower logarithmic score needs to be chosen.

Chapter 2

MSMpred

MSMpred (<https://www.grbio.eu/pubs/MSMpred/>) is a shiny app with two main goals: 1) to fit a MSM from specific data; 2) to predict the clinical evolution for a given individual based on a previously fitted MSM.

As **MSMpred** is mainly designed for clinicians or researchers with little knowledge about MSMs, we have tried to make it very easy to use, to implement all the statistical part in a intuitive way and to include interpretations for the different outputs. Programming skills are not required to use **MSMpred**.

To achieve both goals, fit a MSM and make predictions, **MSMpred** has different sections that make the process quite intuitive. In each of those sections the user works on different aspects of MSMs using parametrizable inputs (e.g., selection of the covariates, characteristics of the new individual).

The main sections of **MSMpred** and their features follow:

- **Home:** it contains a short description of the example dataset, and a brief explanation of the characteristics that the new dataset must have (format of the dataset, names, etc.), as well as some indications of how the app works.
- **Data:** the user uploads his/her own dataset or decides to work with the example dataset (a dataset related with the DIVINE project).
- **Model specification:** the user defines the transitions of the model using buttons, selects the covariates to include in the model, and specifies the follow-up time of the study.
- **Exploring the data:** the app shows some descriptive graphs and tables of the selected covariates, box-plots of the length of stay in each initial/transient state, cumulative incidence and survivals curves for the time until each absorbing state, and non-parametric graphs representing the instantaneous hazards of the transitions over time.
- **Fitted model:** using drop-downs the user decides the type of model to be fitted and receives a summary of the fitted model. For the moment only Markovian Cox models are available.
- **Graphics:** the user receives several forest plots that represent the hazard ratios and confidence intervals of the covariates taken into account in each transition.
- **Model validation:** different graphs related with residuals (e.g., Schoenfeld residuals) that allowed to validate the fitted model are provided.
- **Predictions:** the app returns some predictions for a new individual based on the information provided by the user.

To create this app, we have used `shiny`, `shinyBS`, `shinyWidgets`, `shinydashboard`, `shinydashboardPlus`, `shinyalert`, `shinyMatrix`, and `shinyjs` packages of the R software. For implementing the multistate model we used the `mstate` package and followed the indications given by Wreede [13]. The plots are made using the packages `ggplot2`, `bshazard`, `cmprsk`, `DiagrammeR`, `LoopDetectR`, `survminer`, `pals`, and the tables using `DT` and `summarytools`. Finally, we have used another packages to work with the data as `dplyr`, `stringr`, and `lubridate`.

1. Data

The user can upload a new dataset, provided that it has the required format. For illustrative purposes the app has an example data that comes from the DIVINE project (this project and the associated dataset will be described in the next section).

The dataset to upload needs to be in a csv format and with the following characteristics:

- Columns separated by commas and decimals separators represented by points.
- Time and status variables related to the different states of the model have to be named as x_time and x_status respectively, where x corresponds to the name of each state (e.g. $death_time$ and $death_status$).
- The initial state(s) has/have to be included in the file following the previous naming and having time equal to 0 when the initial state(s) is/are not transient.
- One variable named $inistat$ should be included with the name of the initial state for each individual.
- The variables that are not named as x_time , x_status , id or $inistat$ are considered as baseline covariates, and their name should not include any number or point.

2. Model specification

The first objective of **MSMpred** is to fit a MSM from the data selected by the user. The first step is to define the model, that is, to specify the states and transitions. For that, based on the names of the columns of the dataset named as x_time and x_status the app identifies which are the states (x), and using some buttons and drop-downs of the **model specification** section the user defines the transitions. Basing on that states and transitions the app shows the diagram of the defined model and the number of events for each transition. For these purposes the `create_graph()` function of the `DiagrammeR` package and the `events()` function of the `mstate` are used. In order to make the diagram more understandable the states are plotted in different colors: orange (initial states), blue (transient) and magenta (absorbing).

Despite a MSM allows the inclusion of recursive transitions and loops in the model (e.g. bidirectional transitions), due to some computational problems when using the `mstate` package our app only allows non-recursive transitions and models without loops. The `find_loops()` function of the `LoopDetectR` package, detects if a recursive transition or a loop is tried to be included by the user. If so, a popup indicates that the transition is not allowed. Sometimes those loops or bidirectional transitions play an important roll on the model. In those cases, some tricks can

be used to take into account these loops: 1) include a new state to represent that the individual goes back to the previous state; 2) create some states specifying the number of times that an individual goes into it. For example, if we want to model a recursive illness-death model (FIG. 6), we can use four states (no severe pneumonia, severe pneumonia, severe pneumonia recovery and death) instead of three (no severe pneumonia, severe pneumonia and death) as in FIG. 9. Or if a recursive illness needs to modeled as epilepsy, different states can be defined depending on the epileptic seizure number (first epileptic seizure, second epileptic seizure, and so on).

Due to that limitations, the *number of events for each transition* equals the number of individuals that make each transition, because individuals only make each transition once.

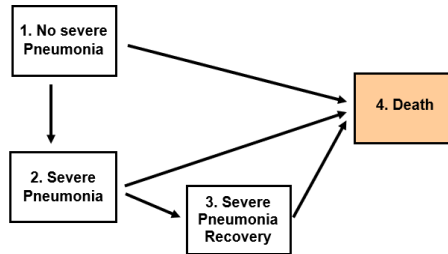


FIG. 9. Recursive illness-death model avoiding recursive transitions

Regarding the selection of the covariates, the user has to decide which covariates wants to include in the model. For the moment, all the covariates have to be measured at baseline, but as future work we would like to implement the feature of adjusting by time-dependent covariates.

Finally, the follow-up time with the pertinent units should be specified in order to delimit the axis related with time in some graphics.

With the abovementioned inputs, the `mstep()` and `expand.covs()` functions of the `mstate` package, allows us to generated a new dataset in long format. This format is characterized by having several rows for each individual, one for each potential transition from the states of his/her path.

3. Exploring the data

In the **exploring the data** section, a descriptive analysis about the selected covariates and the different states of the model is shown. In the **descriptive** subsection, each selected covariate is described by means of a frequency bar chart for factors and by an histogram for numerical covariates. Furthermore, a table with some descriptive statistics is provided: absolute and relative frequencies for each category of the factors and mean, minimum, median, maximum, and interquartile range (IQR) for numeric covariates. The number and percentage of valid values are also returned.

It is important to know how much time an individual remains in a specified state. In the **length of stay** subsection the distribution of length of stay of the individuals in each initial or transient state is shown using box-plots.

In the **time until absorbing state** subsection, the time until these states is analysed by representing the pertinent cumulative incidence and the Kaplan-Meier curves. For a time t , the cumulative incidence indicates the proportion of individuals who go to a specific state, while the survival curve shows the probability of reaching that state, given that the individual has not experienced such state before time t . When data contains censored individuals the cumulative incidence curve of those individuals to describe the time when those individuals are censored. To obtain these curves we applied the `cuminc()` and `survfit()` functions of the `cmprsk` and `survival` packages, respectively.

In the last subsection of this **exploring the data** section, **instantaneous hazards**, the app represents the non-parametric instantaneous hazards of transition over time. To obtain those smoothed estimations of the instantaneous hazards of the transitions we used the function `bshazard()` of the package `bshazard`. In order to reduce the amount of information represented in each graph, a starting and an ending state of interest have to be chosen. Once the app receives these inputs, two graphs are represented showing: (1) the instantaneous hazards of all the potential transitions from the selected starting state; and (2) the instantaneous hazards of all the transitions that go to the ending state of interest. Sometimes if only a few individuals make a specific transition, it is not possible to estimate the instantaneous hazard of that transition. In those cases, the curve of those transition is not plotted.

Those graphics can be stratified by some covariates selected in the **model specification** section. In case of selecting a categorical covariate, a plot is returned for each category, and in case of a numerical covariate, two plots are returned stratifying by the median.

4. Fitted model

To fit the MSM, first of all, it is necessary to decide the type of model. Currently, only Markovian Cox models are available, but in the future, we would like to include other type of models (e.g., non-Markov or parametric models). To fit the model, the function `coxph()` of the `survival` package is used.

In the current version of `MSMpred` if in the **model specification** section some covariates were selected, the same subset of covariates is included for all transitions; otherwise (no covariate was selected), a null model is fitted. Future extension will include the possibility to choose a different subset of covariates for each transition, as this is one of the main advantages of the MSM.

When the Cox model is selected, three tables are returned:

1. **Table of model coefficients.** Each row represents a transition for a given covariate named in the first column of the table as *covar* ($k \rightarrow l$) where *covar* indicates the name of the covariate and $k \rightarrow l$ the transition of interest. For each covariate and transition the estimated coefficient (*coef*), the estimated hazard ratio and its confidence interval (*HR (95%CI)*) and the p-value (*p-value*) are provided. Those values indicate which is the association of the covariate *covar* on the risk for a specific transition.
2. **Table of likelihood.** The values of the log-likelihood of the fitted and the null model. If a null model is fitted, the values of both log-likelihoods match.

3. **Table of goodness of fit.** The outputs of some statistical tests (*test*) of goodness of fit: likelihood ratio; Wald; and score tests. For each test, the value of the statistic, the degrees of freedom (*df*) and the p-value (*p-value*) are reported.

The aim of the tests shown is to contrast the next hypothesis for all the transitions $k \rightarrow l$ and covariates q :

$$\begin{cases} H_0 : \beta_{kl,q} = 0, & \forall q \in \{1, \dots, p\} \\ H_1 : \beta_{kl,q} \neq 0, & \exists q \in \{1, \dots, p\} \end{cases} \quad (24)$$

where $\beta_{kl,q}$ is the regression parameters of the covariate q in the transition from k to l .

This hypothesis is analysed using three different statistics:

- Likelihood ratio statistic:

$$W_{LR} = 2 \log \left(\frac{L(\hat{\beta}_{kl} | \mathbf{Z})}{L(\mathbf{0} | \mathbf{Z})} \right), \quad (25)$$

where β_{kl} is the estimated vector of regression parameters, $\mathbf{Z} = (Z_1, \dots, Z_p)$ is the covariate vector, $R_k(t_{i,kl})$ is the risk set with the individuals at risk in state k at time $t_{i,kl}$ and $L(\hat{\beta}_{kl} | \mathbf{Z}) = \prod_{i=1}^n \frac{\exp(\hat{\beta}_{kl}^T \mathbf{Z}_i)}{\sum_{j \in R_k(t_{i,kl})} \exp(\hat{\beta}_{kl}^T \mathbf{Z}_j)}$ is the partial likelihood of the transition $k \rightarrow l$.

- Wald statistic:

$$W_W = (\hat{\beta}_{kl} - \mathbf{0})^T \mathcal{I}(\mathbf{0} | \mathbf{Z}) (\hat{\beta}_{kl} - \mathbf{0}), \quad (26)$$

being $\mathcal{I}(\beta | \mathbf{Z}) = E\{S(\beta | \mathbf{Z})S(\beta | \mathbf{Z})^T\}$ and $S(\beta | \mathbf{Z}) = (S_1(\beta | \mathbf{Z}), \dots, S_p(\beta | \mathbf{Z}))$, where $S_i(\beta | \mathbf{Z}) = \partial \log L(\beta | \mathbf{Z}) / \partial \beta_i$.

- Score statistic:

$$W_S = S(\hat{\beta}_{kl} | \mathbf{Z})^T \mathcal{I}(\hat{\beta}_{kl} | \mathbf{Z})^{-1} S(\hat{\beta}_{kl} | \mathbf{Z}) \quad (27)$$

Under the null hypothesis, all those statistics follow a χ_p^2 distribution with the degrees of freedom p being the number of coefficients of the model.

Each table has its own interpretation:

1. **Table of model coefficients.** If the covariate is a factor, the hazard ratios compare the hazard rates of each category with a reference category.
 - Positive coefficient. For example, regarding sex, a estimated coefficient, $\hat{\beta}_{kl,M} = 0.56$ implies that a male (M) has $\exp(\hat{\beta}_{kl,M}) = \exp(0.56) = 1.75$ times more risk of transition from the state k to the state l than a female (F, reference category) with the same baseline characteristics.
 - Negative coefficient. For instance, the estimated coefficient is $\hat{\beta}_{kl,M} = -0.1$, a male has $1 - \exp(\hat{\beta}_{kl,M}) = 1 - \exp(-0.1) = 1 - 0.9 = 0.1$ times less risk of transition from the state k to the state l than a female with the same baseline characteristics, or what is the same, the female has $1/\exp(\hat{\beta}_{kl,M}) = 1/0.9 = 1.11$ times more risk of transition than a male in equal conditions.

If the covariate is numeric the hazard ratios, compare two values of this covariate.

- **Positive coefficient.** For example, if we compare two individuals with the same characteristics, but one is 50 years old and the other 65 years old, and the value of the coefficient that we obtain is $\hat{\beta}_{kl,age} = 0.02$, the 65 years old has $\exp((65 - 50) \times \hat{\beta}_{kl,age}) = \exp(15 \times 0.02) = 1.35$ times more risk of transition than the 50 years old.
 - **Negative coefficient.** For instance, if we compare two individuals with the same characteristics, but one is 55 years old and the other 60 years old, and the value of the coefficient that we obtain is $\hat{\beta}_{kl,age} = -0.04$, the 60 years old has $\exp((60 - 55) \times \hat{\beta}_{kl,age}) = \exp(5 \times (-0.04)) = 0.82$ times less risk of transition than the one with 55 years, or what is the same, the one with 55 years has $1/\exp((60 - 55) \times \hat{\beta}_{kl,age}) = 1/\exp(5 \times (-0.04)) = 1/0.82 = 1.22$ times more risk of transition than the 60 years old.
2. **Table of likelihood.** The log likelihood of a model is used to compare the fitting of different models. Models with higher log likelihood provide a better fit to data.
 3. **Table of goodness of fit.** The three tests shown in that table assess the hypothesis that no associations are significant in any transition (i.e., if all the coefficients of the model can be assumed equal to 0). In the column *test* the value of the statistic is reported, in the column *df* the degrees of freedom (equal to the number of estimated coefficients) and the column *p-value* contains the p-value resulting from the corresponding test. Is important to take into account that the p-value is rounded up to the sixth decimal position, so if $p\text{-value} < 10^{-6}$, the app will consider the p-value as 0. Although the three tests could be useful in some situations, for a global test, the likelihood ratio test is preferred over the two other options [14].

As in the **predictions** section we will use this model to make predictions on new individuals, it is necessary to analyse the predictive performance of the model. For that, **MSMpred** allows to compute the logarithmic score clicking on the button *compute the logarithmic score*. It is not computed automatically as it has a high computational cost; its calculation takes more than one minute depending on the included covariates.

To compute the logarithmic score at time of interest (follow-up time selected in the **data specification**), the dataset is randomly split into training (70%) and test (30%) groups. The former is used to fit the model, while the later are used to assess the predictive performance. It is worth to mention that the coefficients of the model fitted using the train dataset could not be the same that those ones resulting from the application of the model on the full data.

When the *compute the logarithmic score* button is clicked a progress bar appears indicating that the computation of the score is being made. The logarithmic score has not a simple interpretation, but it is useful to compare different models, being lower values preferable.

5. Graphics

In order to make a more visual app, in the **graphics** section some forest plots representing the estimated hazard ratios and their 95% confidence intervals are returned. The user needs to choose between a specific transition and he/she receives

the forest plot of the hazard ratios associated to the covariates taken into account on that transition. We say that the covariate has an *effect* on the transition of interest if the confidence interval of these specific covariate and transition does not cover the 1, and that it does not have a significant effect otherwise. If no covariate is selected, a popup is shown indicating that the graphic cannot be shown.

In the case of the numerical covariates the user could decide the difference of units to consider in the computation of the hazard ratio and its 95% confidence interval.

6. Model validation

The assumptions of the model are assessed in the **model validation** section. When a Cox model is fitted three main premises are assumed: 1) linearity of the numerical covariates; 2) absence of influential observations; 3) proportionality of the hazards. For each transition, those assumptions can be evaluated, in a graphical way in the subsections **linearity**, **influential observations** and **proportionality of the hazards**, respectively. When no covariate is selected, a popup appears indicating that the premises cannot be assessed.

In the **linearity** subsection, the user selects the numerical covariate as well as the transition of interest for which the assumption wants to be assess. A graph representing the martingale-based residuals of the selected transition and covariate as function of covariate values is reported. If the smoothed curve along the x axis is reasonably linear, we can assume the linearity of that covariate in that specific transition, otherwise, a transformation of the covariate in that specific transition should be implemented.

In the **influential observations** subsection, we can find a graph representing the dfbetas residuals versus each covariate for a specific transition. These graphs help to detect if there is any potential influential value. Those residuals represent the difference between the estimate obtained when fitting the model considering all the individuals and the one obtained when fitting the model without this particular individual. The individuals that have a dfbetas residual that is far away from the other residuals have a higher influence on the model estimates.

Finally, in the **proportionality of the hazards** subsection the Schoenfeld residuals are used. For each covariate related with a selected transition a graph with those residuals is obtained. Each of these graphs should be independently assessed, and we say that the proportionality of the hazards holds for that specific covariate and transition if the estimated 95% confidence bands entirely cover the horizontal line at $y = 0$.

We used the `ggcoxdiagnostics()` function from the `survminer` package to obtain those residuals and to plot the `ggplot` graphics. This function computes martingale-based, dfbetas or Schoenfeld residuals depending on the `type` argument.

7. Predictions

For the second objective of **MSMpred**, predict the clinical evolution for new individuals, the model previously fitted is used. This prediction can be done for one or two individuals at the same time, making possible the comparison of the evolution of individuals with different profiles.

The predictions are obtained based on the profile of a new individual by means of selecting values of his/her baseline covariates. Also, different aspects of those individuals need to be specified: the current state and the time for which the prediction wants to be done. By default the app makes the prediction for an individual that has the first of the states as current state, and a profile constituted by the first categories of the categorical factors and the medians of the numerical covariates. Those predictions are done over 30 days.

Both a numerical and a graphical output are returned. The numerical output provides the probability of being in each state in the selected time. Those values are computed using the `msfit()` and `probtrans()` functions of the `mstate` package. Additionally, a transition probability plot is represented to have a global vision of those predicted probabilities for any moment before the selected time point. This plot can be represented in a stacked or non-stacked way: in the non-stacked plot, probabilities are represented by curves over time and in the stacked plot, those probabilities are cumulated for each point time by means of coloured shaded areas. Those outputs represent the probability of being in each state after some time, regardless of the intermediate path.

8. Help

The **help** section takes the user to another tab where some indications about the inputs and outputs of the other sections can be found, as well as a guide about how to interpret the obtained outputs. Furthermore, every section and subsection of the app contains a help box that takes the user to the pertinent help tab.

Chapter 3

Case study: DIVINE project

For illustrative purposes, we show how **MSMpred** works using the example dataset from the DIVINE project (2020PANDE00148) that is funded by Generalitat de Catalunya.

The team of this project consist of 9 members from Instituto de Investigación Biomédica de Bellvitge (IDIBELL) and 9 members from Universitat Politècnica de Catalunya (UPC), among those we are 12 biostatisticians and 6 clinicians.

This project has the following four main objectives

- (1) Identify the most clinically relevant prognostic factors for the events,
- (2) **Develop a prediction tool to identify high-risk individuals,**
- (3) Estimate the incubation time period of the SARS-CoV-2,
- (4) Assess the patients' profile over time,

and **MSMpred** have been developed to achieve the second objective of this project.

In order to achieve those objectives, we have data on more than 5,000 hospitalized adult COVID-19 patients from 8 Catalan hospitals during the first five waves of the pandemic. The dataset contains information about relevant events such as death, severe pneumonia and invasive mechanical ventilation (IMV). **MSMpred** contains a subset (n=2048) of this dataset with the information that is needed to fit a MSM, corresponding to the patients without ceiling of care of the first Catalan wave of the pandemic (March-April 2020).

Different models have been designed for the first Catalan pandemic wave, including as states the main outcomes (discharge and death) together with objective interventions during hospitalization such as non-invasive or invasive mechanical ventilation, until ending up with the model that considers 7 states and 14 transitions shown on FIG. 7. Those states do not overlap so individuals can only be in one state at each time. Below you can find a little explanation of each state:

1. **No severe pneumonia** (*nopneum*): patients that are hospitalized due to COVID-19 but do not have severe pneumonia.
2. **Severe pneumonia** (*pneum*): patients that are hospitalized due to COVID-19 and have severe pneumonia.
3. **Severe pneumonia recovery** (*reco*): patients that had severe pneumonia while hospitalized due to COVID-19, they recovered but they are still hospitalized.

4. **Non-invasive mechanical ventilation** (*NIMV*): patients that need non-invasive mechanical ventilation while hospitalized due to COVID-19.
5. **Invasive mechanical ventilation** (*IMV*): patients that need invasive mechanical ventilation while hospitalized due to COVID-19.
6. **Discharge** (*dcharg*): patients that go home or to another hospital after recovering from COVID-19.
7. **Death** (*death*): patients that die in the hospital due to COVID-19.

For that model we have created the state severe pneumonia recovery (*reco*) for recovered patients still hospitalized. This state is necessary because a patient with severe pneumonia (maybe also with non-invasive or invasive mechanical ventilation) needs to be recovered before being discharged. As **MSMpred** does not allow to include loops in the model, this state was created to avoid recursivity. We performed an imputation for the recovery date in those patients that do not need any type of mechanical ventilation because this information was not registered. We assumed that the patient had to be recovered at least 1 or 2 days before the discharge date.

For each state the dataset has variables named *x.time* and *x.status* through which the path of each patient is described. The variable *x.time* contains the time until the state *x* is reached for the first time, and the variable *x.status* indicates if this state is reached or not. If the patient does not reach the state *x*, *x.status* takes value 0 and *x.time* takes the last observed time of the patient. We can interpret that the time until this state *x* is censored at the last observed time as the patient has not reached that state.

FIG. 10 illustrates the path of two specific patients (*id* = 8 and *id* = 1). Both are patients admitted into the hospital without severe pneumonia (*nopneum.time* = 0, *nopneum.status* = 1). The first patient is diagnosed with severe pneumonia after 1 day in hospital (*pneum.time* = 1, *pneum.status* = 1), he/she needs non-invasive mechanical ventilation at day 2 (*NIMV.time* = 2, *NIMV.status* = 1) and invasive mechanical ventilation at day 3 (*IMV.time* = 3, *IMV.status* = 1) and finally he/she dies at day 10 (*death.time* = 10, *death.status* = 1). Consequently, that patient has not reached the states *reco* and *dcharg* and both states are censored at time 10 (*reco.time* = 10, *reco.status* = 0 and *dcharg.time* = 10, *dcharg.status* = 0).

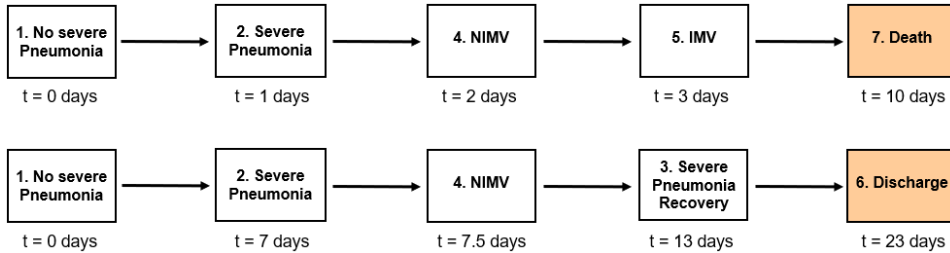


FIG. 10. Model

The second patient has a better evolution since, even if he/she needs non-invasive mechanical ventilation at day 7.5, he/she recovers from severe pneumonia at day 13 and he/she is discharged at day 23.

Despite the time until each state is analysed through the difference between the date of entry in the previous and actual state, in the abovementioned patient we can observe that the time until non-invasive mechanical ventilation is 7.5 days. This is because the dataset has patients that enter in two states the same day, and this

is not allowed in MSMs. To solve this problem, we add a half-day lag for patients that enter in two states at the same day. As the second patient of FIG. 10 was diagnosed with severe pneumonia and needed non-invasive mechanical ventilation the seventh day, we made the half day imputation obtaining $NIMV_time = 7.5$.

It is worth to mention that this data has not censored individuals. This is because instead of ending the follow-up of the patients when ending the study period, we followed collecting the information of the patients until they go to an absorbing state (*death* or *dcharg*). In addition, as we are working with hospitalized patients there is not lost to follow-up.

Apart from the information related to the states, data contains some baseline covariates:

- **Sex (sex)**: dichotomous covariate with categories *Men* and *Women* representing the sex of the patients.
- **Age (age)**: numeric covariate that represents the age of the patients in years.
- **Pneumonia severity index (psi)**: numeric covariate that represents the severity of the pneumonia.
- **Cardiovascular diseases (card_vasc)**: a dichotomous covariate with categories *No* and *Yes* depending on if the patients have any cardiovascular disease or not.
- **blood oxygen saturation/oxygen supply (safi)**: numeric covariate that represents the respiratory limitations in mmHg. The ideal value is $100/0.21 = 476$ mmHg, but as we are working with hospitalized patients they have a lower value.
- **Charlson index (charlson_fact)**: index that predicts 10-year mortality taking values between 0 and 12 and higher values are associated with higher comorbidities. We categorized this index in three categories: $[0,1) \rightarrow low$, $[1,3) \rightarrow mild$ and $> 3 \rightarrow very\ high$.
- **C-reactive protein (crprot)**: a numeric covariate that represents the value of the C-reactive protein (CRP) of each patient in ng/ml.
- **Lymphocytes (lympho)**: a numeric covariate that represents the number of 10^3 lymphocytes per mm^3 of each patient. The range of normal values goes from 1000 cells/ mm^3 to 4.8 cells/ mm^3 .

We now explain the functioning of **MSMpred** using this example data.

1. Data

As we want to work with the example data, in the data section we don't have to upload any file. For confidentiality reasons the app only shows the information of 20 patients, although it internally works with all the individuals of the DIVINE cohort.

2. Model specification

As we are working with the example dataset, it is not necessary to define the transitions, as they are defined automatically. However, one can modify this default transitions. In the *multistate model diagram* box, the diagram of the model shows that there are two orange initial states (*nopneum*, *pneum*), three non-initial

blue transient states (*reco*, *NIMV*, *IMV*) and two magenta absorbing states (*death*, *dcharg*). In the *number of events for each transition* box, we observe that the transitions $reco \rightarrow death$ and $NIMV \rightarrow death$ are only made by 12 and 10 individuals respectively. Due to that, if we choose too many covariates we could have convergence problems that lead to not calculable probabilities when making predictions ([15]).

We select in the *covariate selection* box the covariates *sex*, *age*, *psi* and *card_vasc*.

Finally, in the *time specification* box the follow-up time and the time unit need to be selected. In this data the time is represented in days, and we can analyse the evolution, for example, 30 days later.

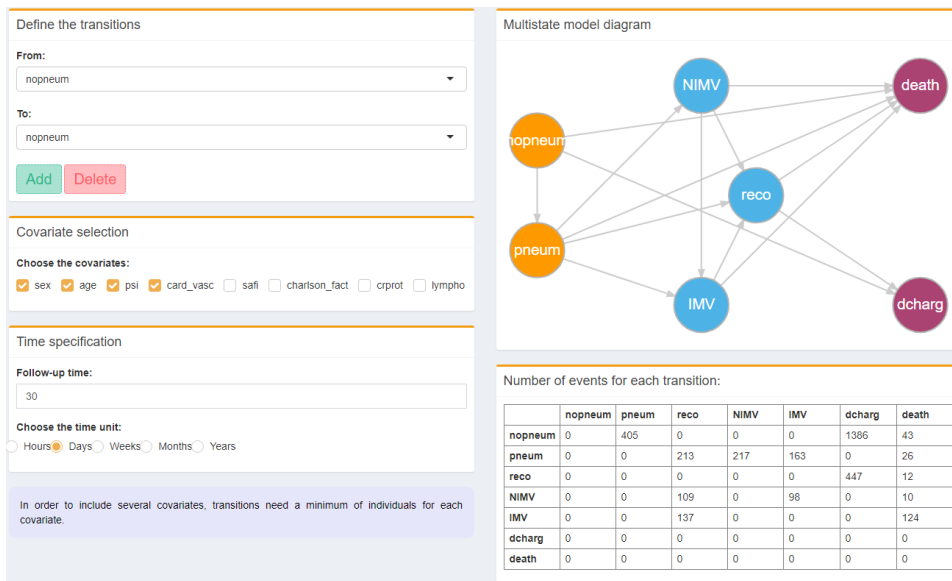


FIG. 11. Model

3. Exploring the data (EDA)

In the **exploring the data** section, descriptive information of the data is shown. In the **descriptive** subsection shown in FIG. 12 we can observe how individuals are distributed through the different categories or values of the selected covariates (*sex*, *age*, *psi* and *card_vasc*). With respect to the age and cardiovascular conditions of the study population, in FIG. 12 we observe that there are more men than women (58.9% vs 41.1%), and that most of them have cardiovascular diseases (79.7%). We can also see that the age range goes from 19 to 96 years with a median and mean age of 59.5 and 58.9 years, respectively. The pneumonia severity index of the patients has a quite large range, as the lowest value is 12 and the highest 184 with a median of 61.

The box-plots for the length of stay are a convenient tool to identify possible outliers before fitting the model. In the box-plots of the **length of stay** subsection shown in FIG. 13, we can observe that the state with the higher median length of stay ($Med = 13$ days) is *IMV*, while the other 4 initial/transient states (*NIMV*, *nopneum*, *pneum*, *reco*) have more similar median length of stays, 3.5, 6, 2 and 4 days, respectively. Observing the upper and lower hinges of the boxes, we also see

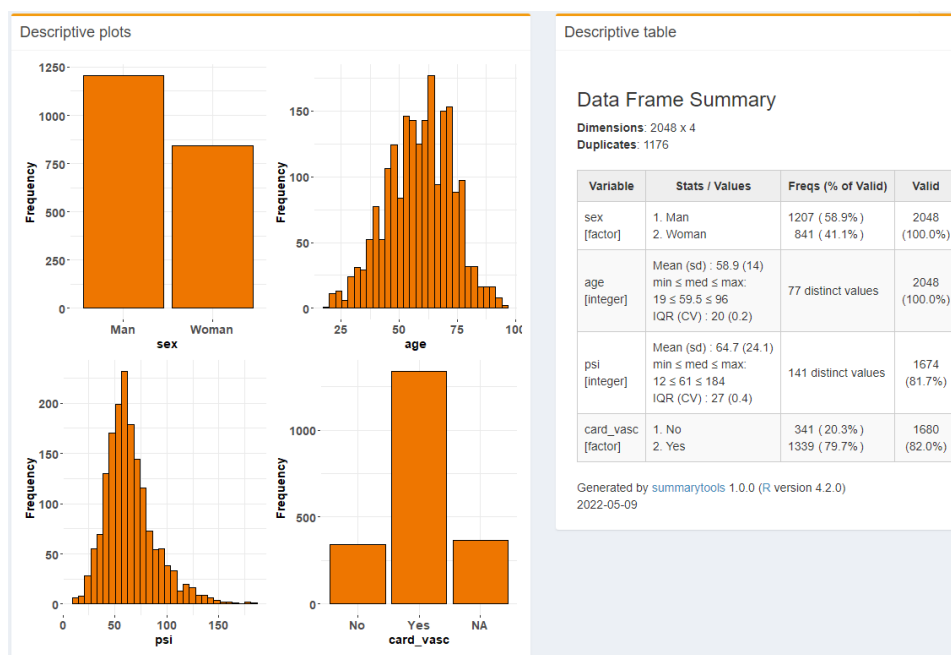


FIG. 12. Exploring the data: covariates

that the number of days in *IMV* are larger than in the other states ($Q1 = 7$ days and $Q3 = 24$ days), followed by the *reco* and *nopneum* with first quartiles 2 and 3 days and third quartile 13 and 9 days respectively, and finally *NIMV* and *pneum* have the lower whiskers that are 1 and 6.5 days more or less. This makes sense since if a patient needs *IMV* is because he/she is in a critical situation so needs a larger time in that state to recover. Regarding to the outliers, there are some patients that deserve more attention (e.g., a patient that has been in *nopneum* more than 100 days).

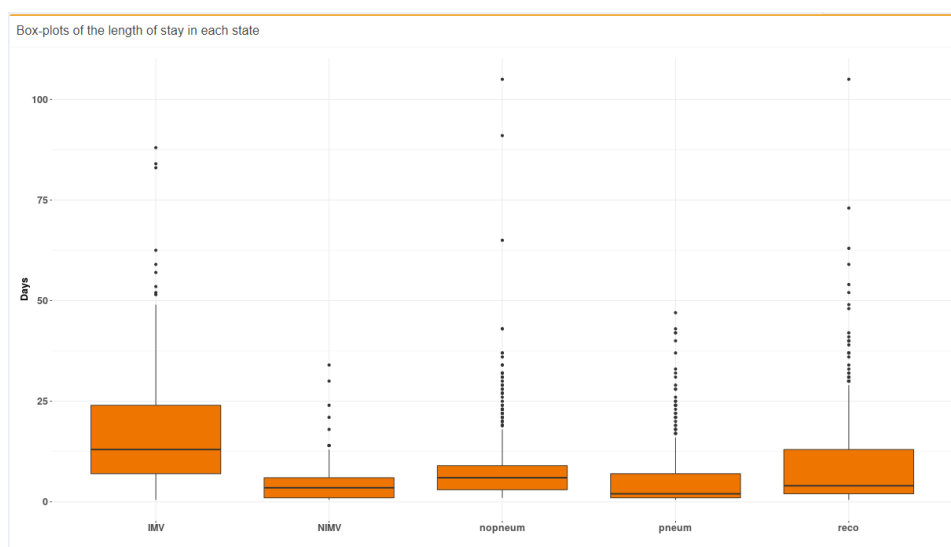


FIG. 13. Exploring the data: length of stay

The cumulative incidence plot of FIG. 14 in the **time until absorbing states** subsection shows that the proportion of patients that die due to COVID-19 during the first 30 days in hospital is more or less the 10% of the sample, while the remaining 90% are discharged. In the plot representing the Kaplan-Meier curves we can see that the probability of reaching the state *death* at day 30 is 0.74 that is higher than the one of reaching the *discharge*, 0.12, state given that the individual has not experienced any of these two absorbing states before time t .

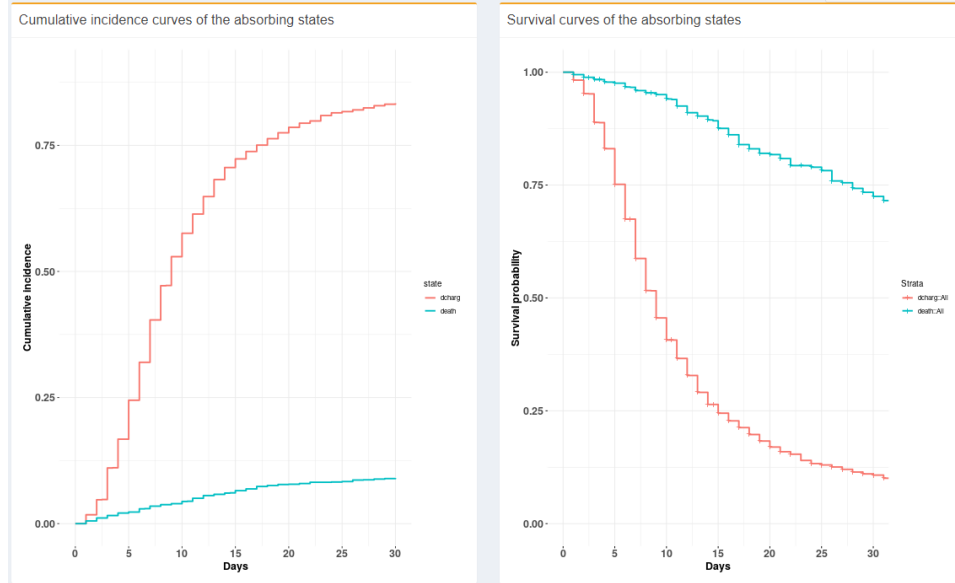


FIG. 14. Exploring the data: time until absorbing states

Finally, in the **instantaneous hazards** subsection a plot like the FIG. 15 is presented. We can stratify the data using one covariate, for example, *sex*. For the transitions starting from the state no severe pneumonia, we can observe that there are not big differences according to the sex. The biggest difference is for the transition *nopneum* \rightarrow *dcharg*, particularly after the seventh day. Regarding the transitions that end in *death*, one can see that there seems to be more differences between the instantaneous hazard of transition between men and women. In general, women that are recovering from severe pneumonia have more risk of dying, while men have more risk of dying when they have non invasive mechanical ventilation.

Two plots representing the instantaneous hazards are obtained if we stratify the data by *age* and select the transitions ending in *death* (FIG. 16): at left, the patients younger than the median age, and at right, otherwise. The number of curves differs between both groups (3 for the younger, and 5 for older group). This is because the transitions from *nopneum* or *NIMV* to *death* do not have enough younger patients and it is not possible to make the estimation of the instantaneous hazard, probably due to those transitions are unusual in young people.

4. Fitted model

In the **fitted model** section, a Time-Homogeneous Markovian Cox model is fitted including all the previously selected covariates in each transition, and three tables are returned giving different information of the model (FIG. 17).

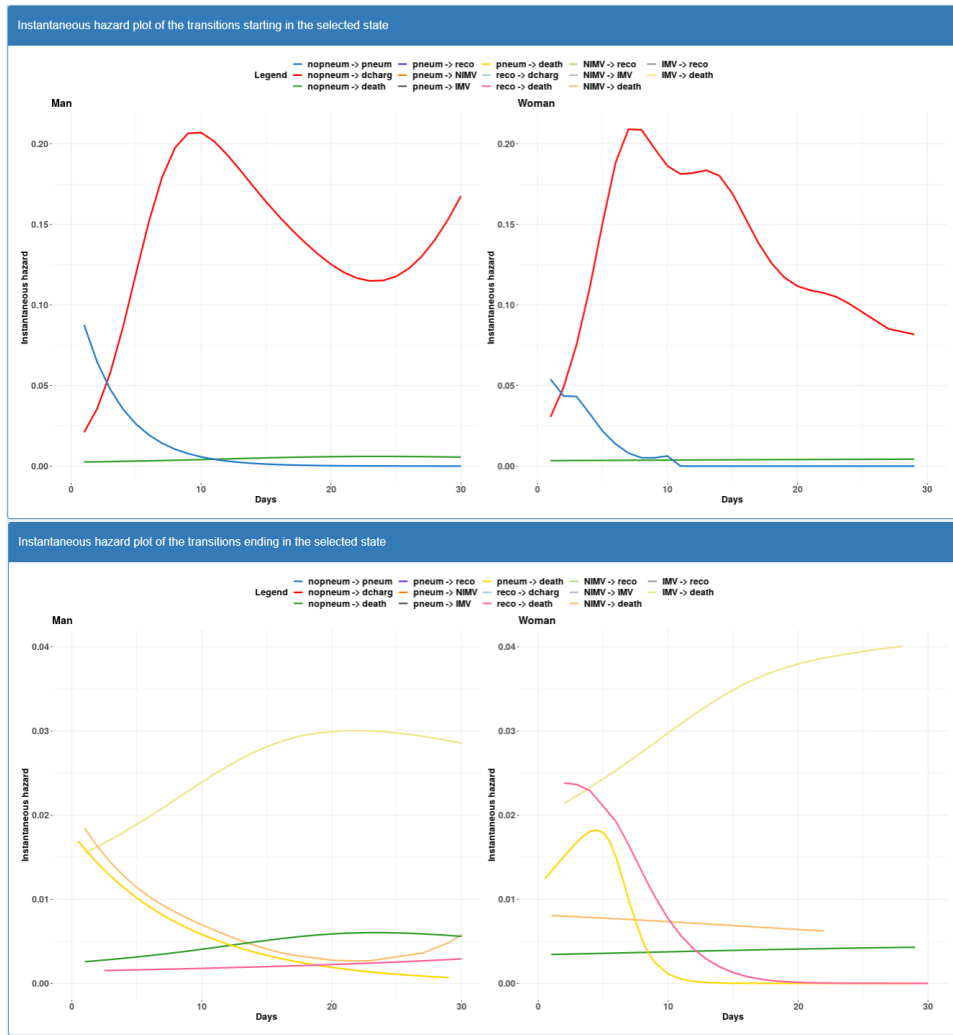


FIG. 15. Exploring the data: instantaneous hazards stratified by sex

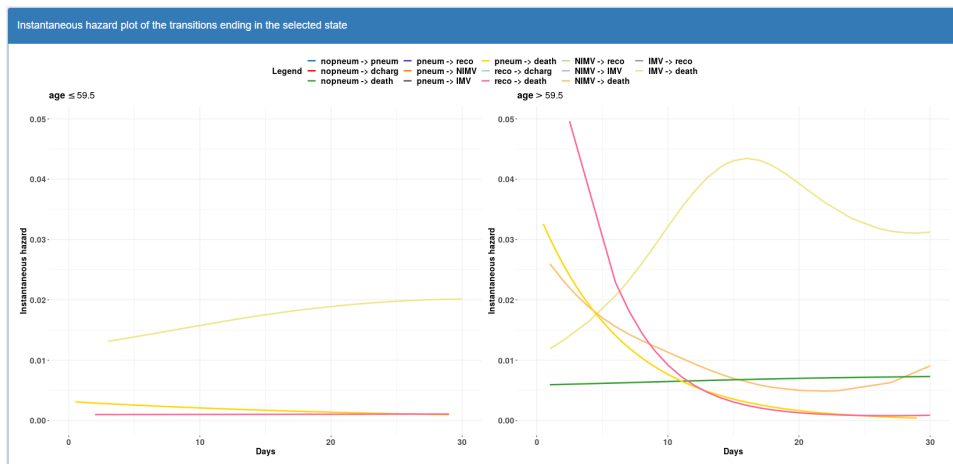


FIG. 16. Exploring the data: instantaneous hazards stratified by age

In the first table, we observe that each of the selected covariates (*sex*, *age*, *psi* and *card_vasc*) has an *effect* in at least one of the transitions. For example, in transition *nopneum* \rightarrow *pneum*, a patient with a cardiovascular disease has 1.48 times more risk of transitioning from no pneumonia to pneumonia than a patient without cardiovascular diseases. In this table it is also important to focus on the effect measure: the estimated hazard ratios. Those hazard ratios and their confidence intervals are very important because in addition to telling if the covariate has an *effect* on the transition or not, they quantify the level of association.

Table of model coefficients:			
Show	10	entries	Search: <input type="text"/>
	coef	HR (95%CI)	p-value
sex (nopneum -> pneum)	-0.259	0.77 (0.61, 0.98)	0.035
sex (nopneum -> dcharg)	0.068	1.07 (0.93, 1.23)	0.336
sex (nopneum -> death)	-0.049	0.95 (0.45, 2.03)	0.898
sex (pneum -> reco)	0.268	1.31 (0.97, 1.77)	0.081
sex (pneum -> NIMV)	-0.239	0.79 (0.55, 1.12)	0.184
sex (pneum -> IMV)	0.028	1.03 (0.71, 1.50)	0.883
sex (pneum -> death)	0.528	1.70 (0.73, 3.94)	0.219
sex (reco -> dcharg)	0.086	1.09 (0.87, 1.37)	0.461
sex (reco -> death)	-1.322	0.27 (0.01, 4.96)	0.375
sex (NIMV -> reco)	0.462	1.59 (0.94, 2.68)	0.083

Showing 1 to 10 of 56 entries

Previous 2 3 4 5 6 Next

Table of likelihood:		
	Null model	Fitted model
log likelihood	-13520.604	-13316.463

Table of goodness of fit:			
	test	df	p-value
Likelihood ratio test	408.282	56	0.000000
Wald test	284.090	56	0.000000
Score (logrank) test	395.345	56	0.000000

FIG. 17. Fitted model

Hereinafter, we will focus on the the transition *IMV* \rightarrow *death*, one of the most important transitions from a clinical point of view. *Age* is the unique covariate that has a relevant *effect* on transitioning from invasive mechanical ventilation to dying. Furthermore, the instantaneous risk of dying increases in older patients. A 70 years old patient with invasive mechanical ventilation has $\exp((70 - 60) \times \text{coef}) =$

$\exp(10 \times 0.032) = 1.37$ times (almost 40%) more risk of dying than a 60 years old patient with the same characteristics.

In the second table of FIG. 17, the value of the log likelihood for the fitted model (-13316.463) and for the null model (-13520.604) are reported. The former will be always higher, but one must resort to the third table to conclude if the difference between two values is important enough.

Based on the obtained values of the tests shown in the third table of FIG. 17 (likelihood ratio test, Wald test and score test) we can reject the null hypothesis ($H_0 : \beta_{kl,q} = 0, \forall k \neq l \in \mathcal{R}, \forall q \in \{1, \dots, p\}$), and assume that in the fitted model at least one of the coefficients of each transition is different from zero.

The value of the logarithmic score, $LS = 0.576$ (FIG. 18) is uninterpretable as we previously mentioned. However, it can be used to compare models. For example, a model with the covariates *sex*, *age*, *psi* and *charlson_fact* has a logarithmic score equal to 0.588 leading to the conclusion that the first model has better predictive performance.

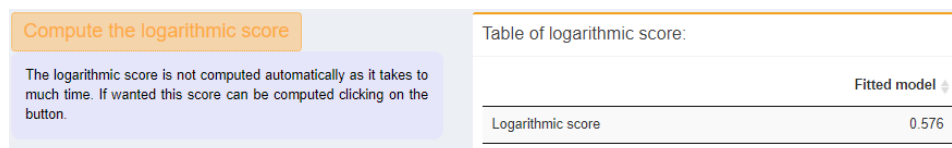


Table of logarithmic score:	
	Fitted model
Logarithmic score	0.576

FIG. 18. Fitted model: logarithmic score

The main limitation of this section is that all the selected covariates are common in all transitions. In the future, we would like to include the possibility of choosing different covariates for each transition. For the moment, to achieve this objective, the following trick can be used: one can fit several models, one for each transition, including the covariates of interest for that specific transition. Then, from each of those models the information related to the transition of interest can be analysed. This can be done because we are working under the Markov assumption, so each transition is estimated independently.

Let's see an example. Suppose that we are interested in fitting the same model as before (FIG. 17), but in the transition $\text{NIMV} \rightarrow \text{death}$ we only want to include the covariates *age* and *sex*. Then, we can fit two models including the covariates: 1) *sex*, *age*, *psi* and *card_vasc*; 2) *age* and *sex*. From the second model we take the information of the transition from non-invasive mechanical ventilation to dying, and the information for the other transitions is taken from the first model.

5. Graphics

The estimated hazard ratios and their confidence intervals are represented by forest plots in the **graphics** section. As the fitted model has too many estimated coefficients, only the hazard ratios of the covariates related with the selected transition are shown (FIG. 19). We can scale the *effect* of the numerical covariates (*age* and *psi*), for instance, representing a change of 10 units instead of 1 unit.

FIG. 19 reveals that *age* is the unique covariate that has an *effect* on transition $\text{IMV} \rightarrow \text{death}$: the risk of transitioning from invasive mechanical ventilation to dying increases 1.38 times when the age of the patient increases 10 years. There

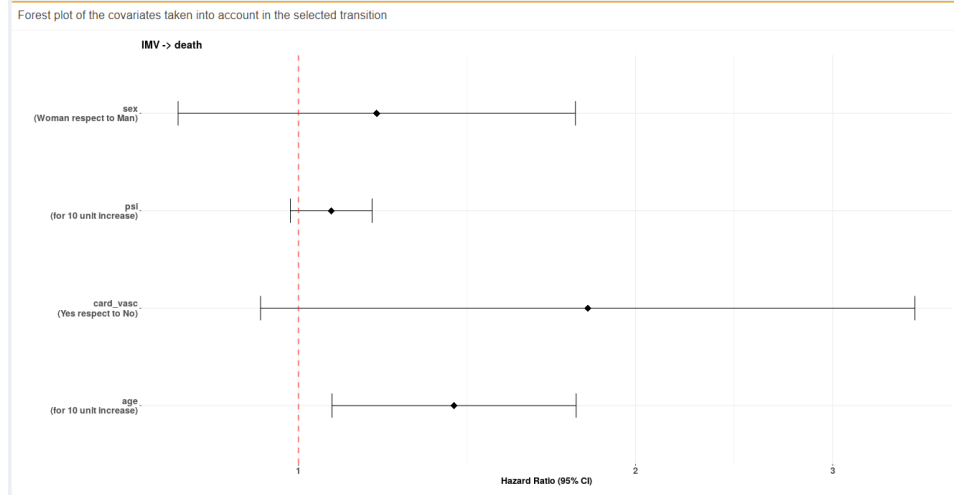


FIG. 19. Graphics

is no evidence to claim that the other covariates (psi , sex and $card_vasc$) have an influence among thus patients that transition from invasive mechanical ventilation to dying.

6. Model validation

The assumptions made when fitting the Cox model should be assessed after every preliminary fitting to reach the final fit. We focus on transition $IMV \rightarrow death$, despite they should be evaluated for all the transitions.

Linearity of the numerical covariates is the first assumption to be assessed using martingale-based residuals. The plot in the left side of FIG. 20 shows those residuals for the age covariate in the transition $IMV \rightarrow death$, while the one in the right side represents the residuals of the covariate psi . In both cases we see that the blue smoothed curve that represents a non-parametric estimate of the trajectory of the points over the covariate values is quite linear, so in this case, the assumption of linearity of age and psi in the transition from invasive mechanical ventilation to dying is sensitive.

The second assumption, absence of influential observations, is analysed in the **influential observations** subsection. For the transition $IMV \rightarrow death$ four plots are shown in FIG. 21, one for each covariate taken into account on that transition: age , $card_vasc$, sex and psi . If we numerically analyse those residuals, the patients that are farther than $2/\sqrt{n}$, being n the number of individuals, need to be considered as influential values [16]. There are 124 patients that go from invasive mechanical ventilation to dying, so we need to consider as influential values all the patients that has a $dfbetas$ residual farther than $2/\sqrt{n} = 2/\sqrt{124} = 0.18$. But those residuals usually are analysed in a graphical way. In FIG. 21 we see that there are not points quite far from the others so, we assume that there are no influential observations.

The last assumption that is the proportionality of the hazards and it is analysed in the **proportionality of the hazards** subsection. In FIG. 22, the Schoenfeld residuals covariates that take part in the transition $IMV \rightarrow death$ are shown. Analysing the smoothed curves and their confidence intervals of the covariates age , $card_vasc$

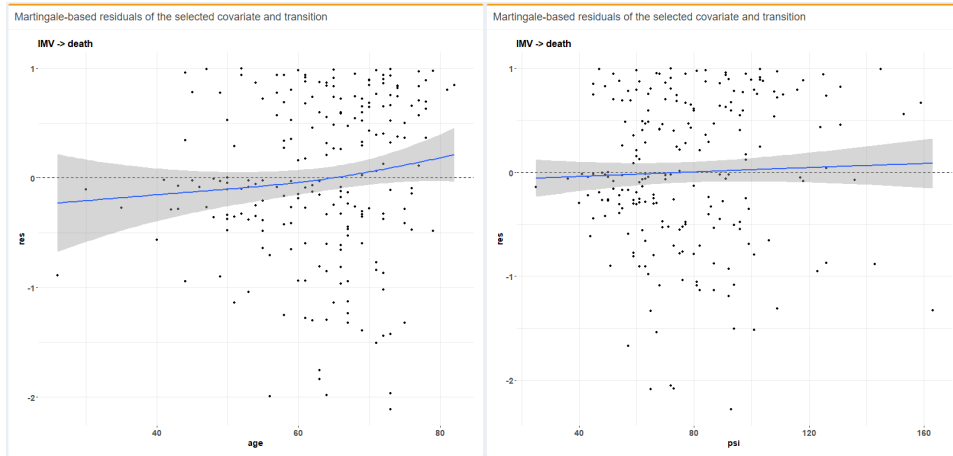


FIG. 20. Model validation: linear assumption for age and psi

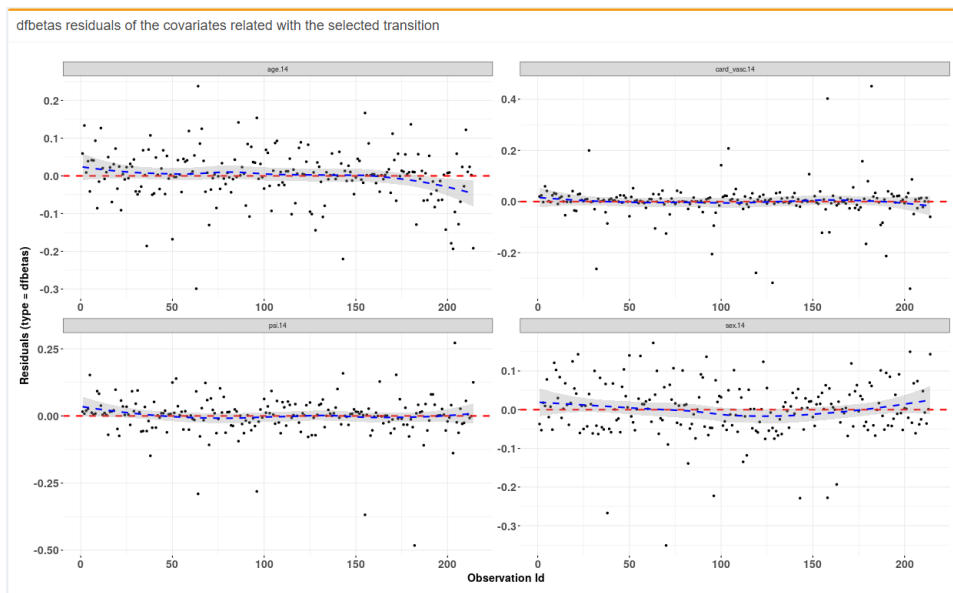


FIG. 21. Model validation: influential observations

and *sex* we see that this premise holds, because the confidence bands cover the line at 0. In the case of the *psi*, there is no full coverage because some part at the end of the horizontal line at 0 is outside of the shaded area. Consequently, we assume reasonable the assumption of the proportionality of the hazards for all the covariates in that transition.

As all the assumptions hold for transition $IMV \rightarrow death$, it is not necessary to go back and redefine the fitting of this transition. But, before drawing conclusions or making predictions based on that model it is necessary to analyse the other transitions.

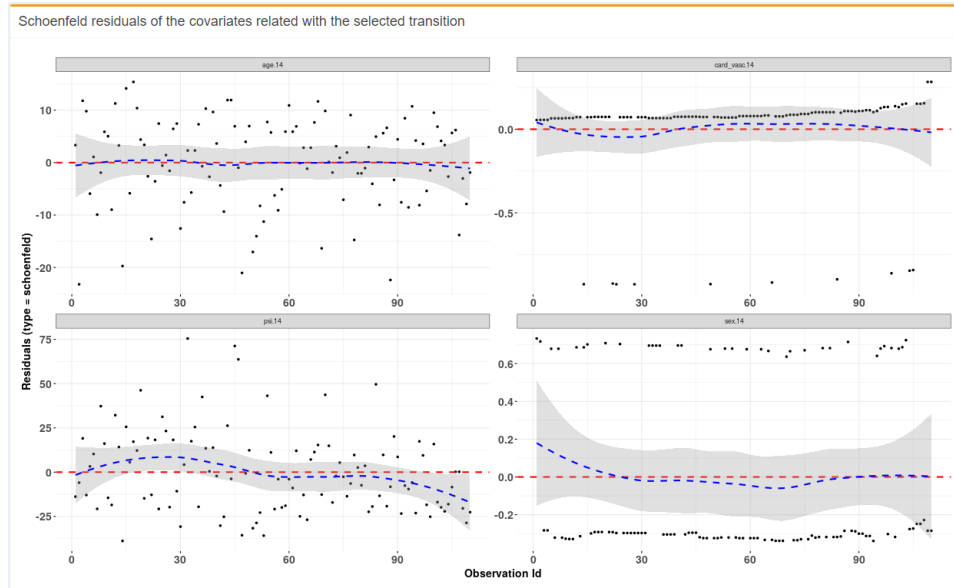


FIG. 22. Model validation: proportional hazards assumpt.

7. Predictions

In the last section of **MSMpred**, **predictions**, some forecast about new patients can be done, allowing to compare the evolution of two patients of different profiles. We are going to compare the evolution at 30 days of two patients that entered into the hospital with severe pneumonia but both of them have completely different characteristics: patient 1 is a young female, that has not any cardiovascular disease and with a pneumonia severity index of 50; patient 2 is an old man, with cardiovascular diseases and with a pneumonia severity index of 110 (FIG. 23).

Characteristics of patient 1	Characteristics of patient 2
Choose the initial state of the new patient: <input type="radio"/> nopneum <input type="radio"/> pneum <input type="radio"/> reco <input type="radio"/> NIMV <input type="radio"/> IMV	Choose the initial state of the new patient: <input type="radio"/> nopneum <input type="radio"/> pneum <input type="radio"/> reco <input type="radio"/> NIMV <input type="radio"/> IMV
sex <input type="radio"/> Man <input type="radio"/> Woman	sex <input type="radio"/> Man <input type="radio"/> Woman
age <input type="text" value="30"/>	age <input type="text" value="80"/>
psi <input type="text" value="50"/>	psi <input type="text" value="110"/>
card_vasc <input type="radio"/> No <input type="radio"/> Yes	card_vasc <input type="radio"/> No <input type="radio"/> Yes
Prediction time: <input type="text" value="30"/>	Prediction time: <input type="text" value="30"/>

FIG. 23. Predictions: characteristics of patients

Once we have defined the profile of the patients, in the *probability of being in each state* box of each patient we can see which is the probability of being in each state after 30 days, no matter which states they have visited before reaching that state (FIG. 24). We can observe that the evolution of patient 1 is very optimistic, 82.9%

probability of having left the hospital before day 30, while patient 2 has a worst prognosis with a probability of 87.1% of dying before day 30.

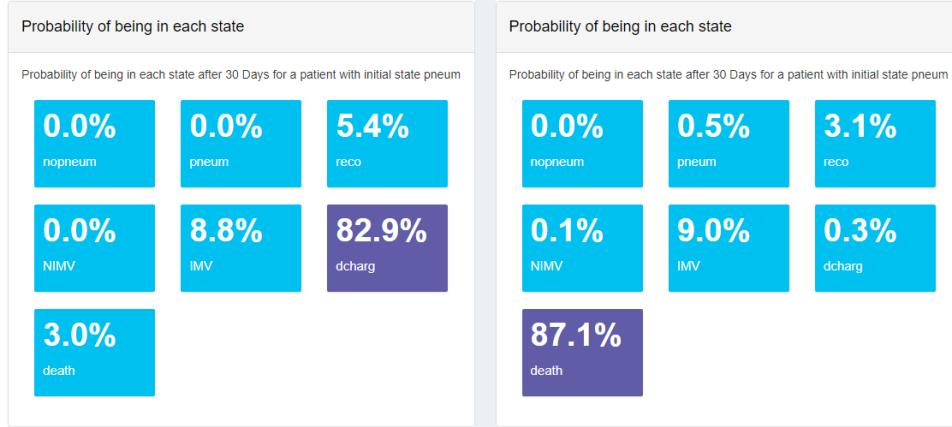


FIG. 24. Predictions: predictions for patients

Finally those probabilities are returned in a graphical way: the transition probability plots are returned in a non-stacked or stacked way. For illustrative purposes in FIG. 25 we show the transition probability plot of patient 1 in a stacked way and the plot of patient 2 in a non-stacked way. Comparing the plots of FIG. 25, the expecting evolution of both patients is completely different. In order to know which is the probability of being in each state in a specific time point, in the case of patient 1 is the height of the color of the state of interest, while in the case of patient 2, it is directly the value of the curve of the state of interest. It is easily seen that the probability of transitioning to *death* for patient 2 is always higher than for patient 1, as well as the probability of transitioning to invasive mechanical ventilation. But, if we analyse the probability of transitioning to discharge the opposite happens, patient 1 has a higher probability than patient 2.

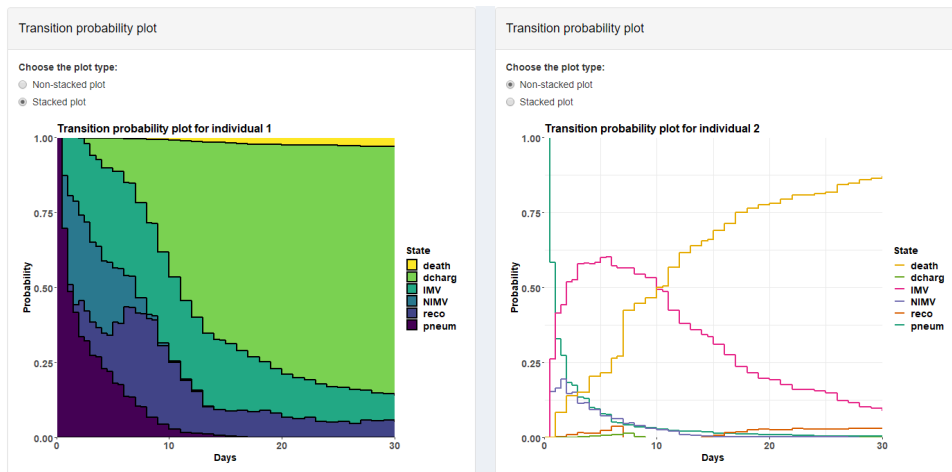


FIG. 25. Predictions: graphical representation for patients

We can clearly see that patient 1 has a good prognosis, because the state with a higher probability after 30 days is discharge which is an optimistic state, while patient 2 has a quite bad prognosis, as *death* is the state with a higher probability after 30 days.

Limitations and future work

We have developed **MSMpred**, an R shiny app that helps clinicians to fit MSM from specific data and make predictions about the clinical evolution of new patients in an interactive way. We have seen that **MSMpred** is an intuitive and visual app and that it is not necessary to have much knowledge of MSM to use it.

We are aware of the limitations that **MSMpred** has and we would like to highlight some of them below.

First, the datasets for MSM are usually in long format, but the app needs wide format and the long format is internally generated by the app.

Second, regarding the covariates included we highlight two main limitations. On one hand, **MSMpred** only allows the inclusion of baseline covariates despite the methodology for MSM allows the inclusion of time-dependent covariates. On the other hand, one of the main advantages of MSM is the use of different covariates for each transition. However, **MSMpred** takes the same selected covariates into account in each transition. This is one of the future improvements that we want to do. Furthermore, if there are few individuals on a specific transition, including too many covariates could lead to convergence problems. Those problems can be detected if some of the estimated coefficients shown in the **fitted model** section have very large absolute values. FIG. 26 shows estimated values for one of the transitions with a low number of individuals in the DIVINE model, transition $NIMV \rightarrow death$. It is clear that the coefficient related with the *sex* has a very low value, -12.356 , and consequently the hazard ratio is almost 0 with an uninformative confidence interval, $(0, \infty)$. Those unexpected values indicate the mentioned convergence problems.

It is possible to solve that problem in two different ways: 1) eliminating covariates for that transition; 2) removing that specific transition. Regarding the former option, it is important to remember that a minimum of events for each covariate is needed. **MSMpred** does not allow to include a covariate just in one transition. Consequently, if we eliminate covariates to avoid the convergence problems, we would obtain a model with very few covariates. A temporary solution is to use the trick explained before: fit several models, one for each transition, including the covariates of interest for that specific transition, and analyse from each of those models the information related to the transition of interest. As **MSMpred** allows to eliminate a concrete transition from the model, it is possible to solve the convergence problems following the second option.

Table of model coefficients:

Show entries Search:

	coef	HR (95%CI)	p-value
sex (NIMV -> death)	-12.356	0.00 (0.00, Inf)	0.975
age (NIMV -> death)	0.091	1.10 (0.97, 1.24)	0.145
psi (NIMV -> death)	0.003	1.00 (0.96, 1.05)	0.886
safi (NIMV -> death)	-0.002	1.00 (0.99, 1.00)	0.558

Showing 1 to 4 of 4 entries (filtered from 56 total entries) Previous Next

FIG. 26. Fitted model: convergence problems

If we make predictions based on a model with convergence problems, we would obtain unfeasible predicted probabilities of being in each state. For instance, with the example dataset, a 90 years old woman hospitalized with severe pneumonia with a pneumonia severity index of 160 and a safi equal to 200, MSMpred provides incoherent probabilities lower than 0 or higher than 1 for some states at 30 days (FIG. 27). In those cases **MSMpred** does not return any result and a popup notifies us that it is not possible to compute those probabilities due to convergence problems.

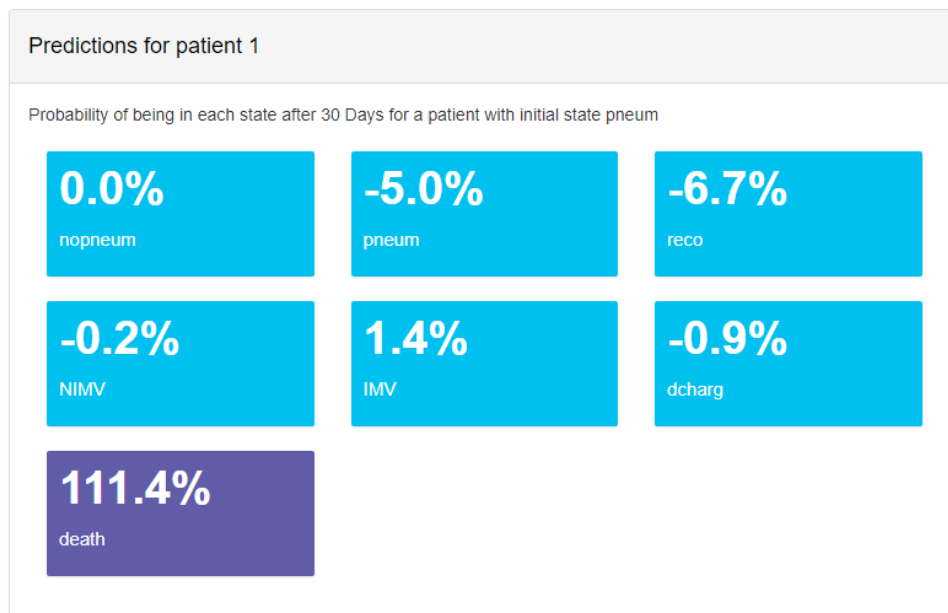


FIG. 27. Predictions: convergence problems

Finally, the estimation of the parameters could be made with several approaches, but our app only implements a semi-parametrical estimation, specifically, in the current version of **MSMpred** only the Cox model is available.

As this is an ongoing app, in the future we would like to improve those limitations. Following we itemize some points that we would like to work in:

- Using some text boxes in the **data** section to allow the user to introduce the name of the states or covariates to be used in the plots. For example, for the covariate named *psi* the user can introduce the name *pneumonia severity index*, or for the state *dcharg* the name *discharge*, and those names will appear in the graphs instead of the ones from the dataset leading to more understandable labels.
- Allowing the models to include not only baseline covariates, but also time-dependent covariates.
- Let the user decide which covariates wants to take into account in each transition of the model.
- In the **fitted model** section, including different type of models such as non-Markov and semi-Markov Cox models, as well as frailty or additive models.
- Possibility of automatically selecting the covariates according to some goodness-of-fit indicator (e.g., AIC). For that, the app internally would make step-wise selection to eliminate the non significant covariates.
- Saving the fitted model in order to allow the comparison of different models.
- In the **predictions** section, compute the median time until entering in each state for the first time.
- When predictions about new individuals are made, **MSMpred** only shows the probabilities of the states that the individuals can reach. For example, if we want to make predictions for a patient that is in *NIMV*, it only shows the boxes with the probabilities of being in *IMV*, *reco*, *dcharg* and *death*, but do not show the boxes of *nopneum* and *pneum* as the patient cannot go back to these states.
- Allowing the comparison of the evolution of more than two individuals. This can be done showing in a table the probabilities obtained when predicting the evolution of new individuals.
- Including a download button to generate a report wit the provided information (e.g., plots, tables). It would be done for each output one by one or via **R Markdown** generating a report with all the information.

References

- [1] Wynants L, Van Calster B, Collins GS, et al. (2020) *Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal*. BMJ. 369:m1328.
- [2] Ursino M, Dupuis C, Buetti N, et al. (2021) *Multistate Modeling of COVID-19 Patients Using a Large Multicentric Prospective Cohort of Critically Ill Patients*. Journal of Clinical Medicine. 10(3): 544.
- [3] Mody A, Lyons PG, Vazquez Guillamet C, et al. (2021) *The Clinical Course of Coronavirus Disease 2019 in a US Hospital System: A Multistate Analysis*. American Journal of Epidemiology. 190(4): 539-552.
- [4] Deschepper M, Eeckloo K, Malfait S et al. (2021) *Prediction of hospital bed capacity during the COVID-19 pandemic*. BMC Health Services Research. 21, 468.
- [5] MSM shiny: <https://stulacy.shinyapps.io/msm-shiny/>
- [6] MSMplus shiny: <https://nskiostatistics.shinyapps.io/MSMplus/>
- [7] MSD app: <https://ph-ivshiny.iowa.uiowa.edu/rpterson/MSDshiny/>
- [8] Unemployment MSM app: <https://johnng.shinyapps.io/Unemployment-MultiStateModel/>
- [9] Meira-Machado L, de Uña-Alvarez J, Cadarso-Suárez C, Andersen PK. (2009) *Multi-state models for the analysis of time-to-event data*. Statistical Methods in Medical Research. 18(2):195-222.
- [10] Cook RJ, Lawless JF (2018) *Multistate Models for the Analysis of Life History Data*, Chapman and Hall/CRC.
- [11] Li J and Valliant R (2011) *Linear Regression Influence Diagnostics for Unclustered Survey Data*. Journal of Official Statistics. 27(1): 99-119
- [12] Crowther MJ, Lambert PC (2017) *Parametric multistate survival models: Flexible modelling allowing transition-specific distributions with application to estimating clinically useful measures of effect differences*. Statistics in Medicine. 36(29): 4719-4742.
- [13] de Wreede LC, Fiocco M and Putter H (2011) *mstate: An R Package for the Analysis of Competing Risks and Multi-State Models*. Journal of Statistical Software. 38(7).
- [14] Harrell FE (2001) *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis*. Springer.
- [15] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996) *A simulation study of the number of events per variable in logistic regression analysis*. Journal of clinical epidemiology. 49(12): 1373-1379.
- [16] Belsley DA, Kuh E, and Welsch RE (1980) *Regression Diagnostics: Identifying Influential Data and Source of Collinearity*. John Wiley.