



UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH



UNIVERSITAT DE
BARCELONA



UNIVERSITAT
ROVIRA I VIRGILI

FACULTAT D'INFORMÀTICA DE BARCELONA (FIB)
FACULTAT DE MATEMÀTIQUES (UB)
ESCOLA TÈCNICA SUPERIOR D'ENGINYERIA (URV)

Factors associated to mortality in patients of the first wave infected by COVID-19 in Spain

Master Thesis

Author:

Marta Barroso Isidoro

Supervisor:

Dario Garcia Gasulla, Barcelona Supercomputing Center

Barcelona, 21 June 2021

Abstract

An ongoing outbreak of coronavirus disease 2019 (COVID-19) emerged in Wuhan since December 2019 and spread globally. Although since then the situation has improved with respect to the number of initial deaths in Europe, there is still a lack of knowledge about the behavior of the disease. In the light of facts, we aimed to analyze the risk factors involved in hospital mortality, and develop methods to predict it using statistical analysis and survival methods.

We selected data retrospectively regarding 1,140 critically ill adult patients with laboratory-confirmed COVID-19 from 63 hospitals in Spain through February 1 to July 31, 2020. Demographic data, symptoms, laboratory values, comorbidities, treatments, and clinical outcomes were collected. The primary outcome was in-hospital mortality. Data were compared between survivors and non-survivors.

Of the 1,140 patients included in the study, the median age was 65 years (IQR 56–71), and 816 (71.58%) were male. Among these patients, 443 (38.86%) died, 585 (51.32%) were discharged, 74 (6.49%) were transferred to health centers and 38 (3.33%) were transferred to other hospitals. Important differences between survivors and non-survivors are observed for chronic kidney disease, heart chronic disease and pulmonary chronic disease.

Survival analysis has comprised two ways of proceeding, approach A models the event of hospital death and approach B models ICU discharge. Standard-Cox in approach A (C-index=0.84) showed that CRP (ICU, 3rd ICU day), SOFA score (3rd ICU day), presence of symptoms and lymphocytes (ICU) are associated with a higher risk of death, and platelets (ICU), glucose (ICU), lactate (ICU) to a lesser extent. For approach B, subdistribution hazard model shows as risk factors: age, glucose (3rd day), lymphocytes (ICU), acute kidney failure, neuromuscular blockers requirement and hemodynamic SOFA (3rd day). In addition, the risk of dying is higher for patients who spent the same time in ICU and IMV (*equal_ICU_IMV_times*). As variables associated with a good prognosis we find time spent in ICU, time spent in IMV, paFi (day3) and leukocytes (day3) for StandardCox and the use of corticosteroids and

platelets (3rd day) for subdistribution hazard model.

Acknowledgements

We would like to thank CIBERES team in the context of BSC for contributing to the successful development of this project and more broadly CIBERES-UCI-COVID. Personally, I would like to thank Raquel Pérez Arnal and Adrian Tormos for their rigorous work highlighting the importance of the good practices that Raquel always instills in us.

Emphasize also the participation of Dario Garcia Gasulla and Sergio Álvarez-Napagao as supervisors and guides since the beginning of the project that motivate us to discuss and reflect on all kinds of aspects, implementation, security, ethics, thus helping to complete a successful work. I express my gratitude to Dario again for supervising the present document with such a level of detail.

Additionally, thanks to Barcelona Supercomputing Center and Professor Ulises Cortés, head of the High-Performance Artificial Intelligence research group, for trusting in my abilities and giving me the opportunity to work with this group.

In the clinical context, we have to thank CIBERES-UCI-COVID for creating and providing us access to the anonymized clinical data set in REDCap platform that lead after to the work performed by HPAI, our research group at BSC. We would also like to Anna Motos, Adrian Cecato y Albert Gabarrus for the help provided in the collection, validation and interpretation of laboratory and ventilatory data collection.

Finally, we thank all the doctors, nurses and clinical scientists who worked in the hospitals during the period of patient recruitment as well as the patients who were involved in this study.

Contents

Contents	vii
1 Introduction	1
2 Related work	3
3 Data and requirements	5
4 Analysis and learning methods	11
5 Statistical Analysis	28
6 Survival Analysis	32
7 Results	49
8 Conclusions	83
9 Future work	87
10 Appendix	89
Bibliography	175

Chapter 1

Introduction

This project aims to carrying out an in-depth, retrospective and multicenter analysis on the distribution, correlations, missing values and survival of covid-infected patients in Spain. Artificial intelligence (AI) has been used for extracting information about the factors involved in mortality, for classifying patients according to certain patterns, and for estimating the time for a group of individuals to experience an event of interest (*e.g.*, reach a critical condition or require mechanical ventilation).

In May 2020, the CIBERES-UCI-COVID [1] project was awarded, funded by ISCIII. CIBERESUCICOVID project gathers data from 69 different Spanish ICUs, including several specific sources such as Getafe hospitals and the SEMICYUC consortium, within the period from February 1, 2020 to June 2021. Note this is the largest data gathering effort for ICU data in Spain. For gathering, processing and exploiting such amount of information, the collaboration of experts from interdisciplinary fields is required. For this reason the CIBERES-UCI-COVID consortium is composed by medical doctors, bioinformatics and AI researchers.

The author of this paper have the latter role and belongs to HPAI research group at BSC. BSC was in charge of developing a complete and unified database to store patient data, perform pre-processing, implemented the required automation process tools to generate scientist reports and carry out statistical and survival analyses. This thesis takes place in this context, focusing and including a significant part on the contributions made by BSC and CIBERES-UCI-COVID consortium.

Chapter 2

Related work

The pandemic caused by the novel COVID-19, has become one of the biggest health challenges worldwide. The SARS-CoV-2 virus is the seventh known coronavirus to infect humans, its emergence makes it the third in recent years to cause widespread infectious disease following the viruses responsible for SARS and MERS. Common human coronaviruses typically cause mild symptoms such as a cough or a cold, but the novel coronavirus SARS-CoV-2 has led to more severe respiratory illnesses and deaths worldwide.

Taking into account the registry of June 20, 2021, there are nearly 178 million confirmed cases and more than 3,86 thousand deaths. The countries that register the highest cases are United States (33,537,95 cases) followed by India (29,881,772 cases), Brazil (17,883,750 cases) and France (5,817,272 cases). However, if we take into account the percentage of deaths, we find Yemen (19.7% death rate), Peru (9.4 % death rate) and Mexico (9.3 % death rate). This makes the COVID-19 pandemic one of the public health problems with a meaningful impact in the history of humanity since the appearance of the last pandemics Influenza A (2009), HIV / AIDS (1980) and “Asian” flu (1957-1958)

Efforts from a research point of view, apart from being directed in the first place to the search for a cure, have focused on characterizing patients with COVID-19 in different phases of the disease (commonly at hospital admission), analyzing risk factors and implement mortality risk calculators or mortality predictors.

Initial studies focus on describing the clinical characteristics and outcomes of critically ill patients with COVID-19. One of the first published studies was [2]. They report clinical characteristics and laboratory findings on the first 799 people with the disease admitted to the isolation ward of a hospital in Wuhan. It was observed that non-survivor patients were on average 17

years older (with no deaths among those aged under 40 and 16.8% of deaths among those aged 40-60), more likely to be male, and more likely to have a comorbidity such as hypertension, diabetes, cardiovascular disease, or chronic lung disease. These results are largely consistent with European studies [3], [4], [5]. The methods used include statistical analyzes comparing survivors and non-survivors. Other analyzes are carried out with more specific cohorts such as patients suffering from a particular disease. For instance, we find [6] that use a cohort of patients with hematological malignancies and [7] patients with lymphoma.

Patients with older age, hypertension, and high lactate dehydrogenase are considered factors that increase the risk of severe disease [8]. These results are similar to many other case series from China [9], [10], [11],[12]. Apart from older age, in [2], the authors reported malignancies, high APACHE II score, high D-dimer level, low paFi level, high creatinine, high hscTnI (high-sensitivity troponin-I) and low albumin level as independent risk factors for mortality. The presence of comorbidities, diabetes, obesity, and smoking are also reported as factors that increase the risk of severe disease in some studies [13], [14]. In particular, in Spain cardiovascular and renal condition appear as risk factors as well [15]. It can be observed that characteristics of COVID-19 may vary depending on the demographic and epidemiological profiles of each country. Commonly used methods seen in the literature comprise univariate and multivariable Cox proportional hazards regression analysis along with Kaplan-Meier curves.

To support decision making and logistical planning in healthcare systems, there is a growing interest in developing machine learning models to predict prognosis and more specifically mortality. In [16], they predict mortality risk using XGBoost algorithm over 3,062 patients with a population of non-survivors of 26.84%. They identified increased age, decreased oxygen saturation ($\leq 93\%$), elevated levels of C-reactive protein ($\geq 130mg/L$), blood urea nitrogen ($\geq 18mg/dL$), and blood creatinine ($\geq 1.2mg/dL$) as primary risk factors. Another Chinese study revealed that lactic dehydrogenase (LDH), lymphocyte and high-sensitivity C-reactive protein (hs-CRP) seem to play a crucial role in distinguishing the vast majority of cases that require immediate medical attention [17].

Chapter 3

Data and requirements

This chapter includes all details regarding the design of the study, data collection and the project requirements and design taking into account the context. In the first place we discuss the details that concern the framework in which the project is developed as well as the variables that are collected 3.1. Later in 3.2 we review the tools that have been used for data collection. Finally in 3.3, we briefly discuss the design of the infrastructure, security aspects and the process of data cleaning and filtering.

3.1 Study design

The present study aims to carry out a retrospective multicenter study to characterize patients admitted to the ICU during the first wave of COVID-19, analyze the risk factors involved in hospital mortality, and develop methods to predict it. Secondly, we have also included the pre-processing analysis comprising the analysis of missing values, outliers, correlations and feature selection that was done in the context of CIBERES-ICU-COVID.

For the study we selected those patients who required invasive mechanical ventilation during the first day of admission to the ICU and who remained ventilated 3 days later. This interest is motivated by the absence of published studies analyzing the influence of laboratory and ventilatory variables on the third ICU day. In addition, information about baseline (*i.e.*, symptoms, comorbidities, previous medication, etc), outcome and gender must be available. This information is required in order to avoid completely unfilled patients. However, it is assumed that we may have missing data for the rest of variables. This is not an issue since imputation can be attempted.

Patients were excluded if they had non-confirmed COVID-19, no data at

baseline or at hospital discharge, or if they were admitted to ICU for other reasons. In that case 64 patients were omitted. For these patients, hospital discharge is motivated by the transfer to social-health centers and other hospitals, and we cannot ensure whether this is due to improvement or deterioration in the evolution of the disease. Thus, the event of hospital death for these patients is unknown.

3.2 Data collection

Patients were enrolled if they fulfilled at least the following criteria: ≥ 18 years old, admission to ICU and laboratory-confirmed diagnosis of COVID-19. These data was anonymized, collected and stored via the REDCap [18] form-based tool, hosted at the Centro de Investigación Biomédica en Red (CIBER), Spain. REDCap is a web-based database for medical and biomedical research support created by the REDCap Consortium [19]. Given the familiarity of the medical partners with this technology, the CIBERES-UCI-COVID project uses this database as single source of truth, which is filled by specialised *data entries* hired by each hospital. Notice the data is collected and introduced into the platform manually by these data entries. The considerable volume of patients that Spanish hospitals have received since the onset of the pandemic makes data collection a complex task and susceptible to errors.

REDCap is structured by entries that in turn contain a set of forms linked to a patient. Although the use of forms can facilitate the registration of variables to medical professionals, they are nevertheless complex to process since they do not follow a linear structure, there are fields that can have several values or that may depend on others fields (*i.e.*, duration of a treatment only appears if a treatment is applied). The data is downloaded through calls to the REDCap API. Since the API does not allow to return more than 1,000 entries at once, we request the data by chunks of N size being N the number of unique patients divided the number of cores of the server. These downloads, known as data migrations, are run periodically to keep our database up to date. Given the number of patients and the amount of data for each one, this process has been parallelized to reduce execution time. This work was contributed significantly by other members of HPAI research group as previous work to set up and enable the rest of the contributions presented in this document.

The rigidity of this platform in the light of executing data analysis, as well as the need for data pre-processing (cleaning and improving the data) led to the creation of a complete and unified database derived from REDCap that would constitute a comfortable framework to work with. This was then one of

the first goals within CIBERES-UCI-COVID project and is further discussed in 4.

Clinical features After the patient has been enrolled to the hospital, previous epidemiological data were collected, including demographic data, comorbidities, clinical symptoms, disease chronology, and treatment administered at hospital admission. Among the collected data we find included vital signs, respiratory support devices (*i.e.*, oxygen mask, cannular high nasal flow, and non-invasive and invasive mechanical ventilation), the use of complementary therapies (*i.e.*, neuromuscular block, prone position and maneuvers recruitment), laboratory findings, arterial blood gases, and mechanical ventilation settings.

Some of these variables contain different measurements, collected in each of the phases that a patient may go through: ICU admission, start of mechanical ventilation, 72–96 h after ICU admission, weaning, ICU discharge and hospital discharge. In particular, for the ICU event, information about hemodynamic parameters and organ dysfunction is also stored, such as the Sequential Organ Assessment Failure Score (SOFA), an important scale for assessing patient severity. Specific data regarding mechanical ventilation since the start of intubation, as well as, at day 3 have been analyzed. Mechanical parameters related to ventilation-induced lung injury (VILI) included tidal volume, respiratory rate, end-inspiratory plateau and peak inspiratory pressures, positive end-expiratory pressure (PEEP), driving pressure, and static compliance of the respiratory system (Crs).

3.3 Global technical requirements and design

In this section, the technical requirements that have conditioned CIBERES-UCI-COVID are presented. In essence, these are related to the creation of a complete, unified and reliable database, the implementation of data pre-processing and filtering tools, and the reporting of results. The database and the pre-processing pipeline will be implemented for the benefit of posterior analysis, including the survival analysis conducted and presented in this work. The latter filtering and reporting of results will be implemented for the benefit of other partners in the project.

Database implementation As mentioned above, one of the main objectives of the project is to generate a database that could be used for any data

analysis or AI purpose. The different entities of REDCap (*i.e.*, comorbidities, previous medication, complications, etc) are represented in different tables where we find one patient per row. For those variables that are taken periodically (*i.e.*, tests), several rows are added in addition with time bound. As database model, a relational database using MariaDB [20] has been implemented. MariaDB comes with a wide range of safety measures and it is faster and more efficient than MySQL.

Data cleaning Missing values and outliers need to be addressed to avoid problems in statistical analyses. We were provided with a list of laboratory and ventilation variables and their normal ranges. For those observations that have variables outside its range, instead of removing the complete observation we ignore these variables and later we treat them as missing values. For further information about missing values visit 4.1.4.

Data filtering From clinical point of view, there is interest in filtering patients according to different factors. The system must be flexible and scalable. It must be possible both to select a set of filters and to incorporate new filters in the future. The following list shows the list of filters implemented in this project:

- Filter initial population using a list of patient ids.
- Filters on filled in variables such as outcome and gender.
- Filters based on variable values (treatments, comorbidities, severity, events, etc.)
- Exclusion filters by id. It is the case of those patients that are not analyzable explained in 3.1.

Report generation The results must be presented in a clear and organized way to facilitate their understanding. For that, we decided that using reports would be the most appropriate way. Indeed, to avoid manually generating a large quantity of very similar reports, those reports need to be generated in a automated way. Thus, given the role it plays in the interaction with the medical team, the generation of reports is of special interest.

The system can generate several types of reports: tables (csv files), correlation reports (matplotlib plots) and text reports (docx files). The latter is used to display results of the different analyses such as missing values and

outlier analyses. It implements methods for description analysis as well. Description analysis distinguishes between numerical and categorical variables. For the first, it is able to generate boxplots, histograms and textual explanations about central tendency measures. For categorical variables, it generates frequency tables and bar plots.

The process for generating a report is composed of three main steps. In the first place, the user instantiates a *Configuration* object. This class contains the minimum variables required to generate reports such as the list of variables and population filters but also contemplates the possibility of adding new fields easily. As a result, the system builds a validated dataframe with the requested variables and the restrictions already applied to it. The conducted analyses are applied to these data and their results are saved in the corresponding report.

Security concerns Finally, considering the personal nature of all data being handled, special focus has to be put onto security measures. All data is stored in a private isolated – via virtualisation – server placed in the EU (Spain). Only an automated GitLab CI/CD pipeline has access to this server unless emergency access is needed, and in this case only authorised researchers can access via SSH tunnelling using a private key. All private data is stored on a MariaDB database. If data has to be uploaded or downloaded by third parties, we use an encrypted – at rest and in transport – distributed S3-compatible storage server. All credentials related to the project are stored in a secure Hashicorp vault. Tokens to access this vault are only available to specific code repositories and authorised researchers. By combining repository-based authentication with a token-based vault, we enforce that the storage and processing of data is carried out, in an automated fashion, by code reviewed and pushed by the authorised researchers.

Target infrastructure The project data flow has been defined in accordance with previously identified technical requirements. First, we seek to have a single, reliable source of truth. This means that all the information comes from REDCap. The other data sources were previously integrated with REDCap. During the data analysis process, we have permission to modify that data in our database. The data is cleaned and examined, both improvements and errors are communicated to medical professionals. They then decide whether to incorporate these modifications to the source of truth.

This results in the scheme shown in Figure 3.1. The two external sources of data from SEMICYUC consortium and Getafe hospitals have a specific parser, which makes them integrable into REDCap. This component also has

another parser, which transforms it from and into our own exploitable database (*HPAI DB* in the diagram). Finally, from this database, several data formats are exported for enabling the posterior processing.

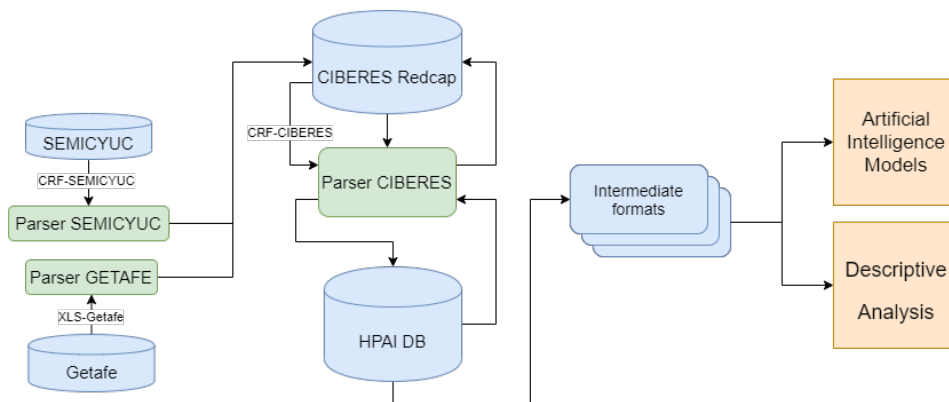


Figure 3.1: Dataflow implemented in CIBERES-UCI-COVID. As cylinders, data sources. In green, data parsers implemented by the authors.

Chapter 4

Analysis and learning methods

This chapter includes all details regarding the data processes implemented as a first step towards learning, as well as the methodologies used later on for experimentation in Chapter 7. First we review and discuss the several pre-processing steps necessary for preparing the data for any posterior analysis in 4.1. Then in chapter 5 we perform some statistical analysis to gain some preliminary insight into the behavior of variables, which can be useful in deciding how to deal with them. Finally, in chapter 6 we discuss the different models we will be using for our particular machine learning goal, one based on survival analysis in patients.

4.1 Data pre-processing

In this section we review the main pre-processing steps performed, prior to any consequent analysis. As depicted in 4.1, the pre-processing is formed by the data collection and filtering process, the definition and analysis of outliers, the analysis of correlated data, a brief descriptive analysis to summarize and see a first shape of population, the detection and handling of missing values and the creation of new features.

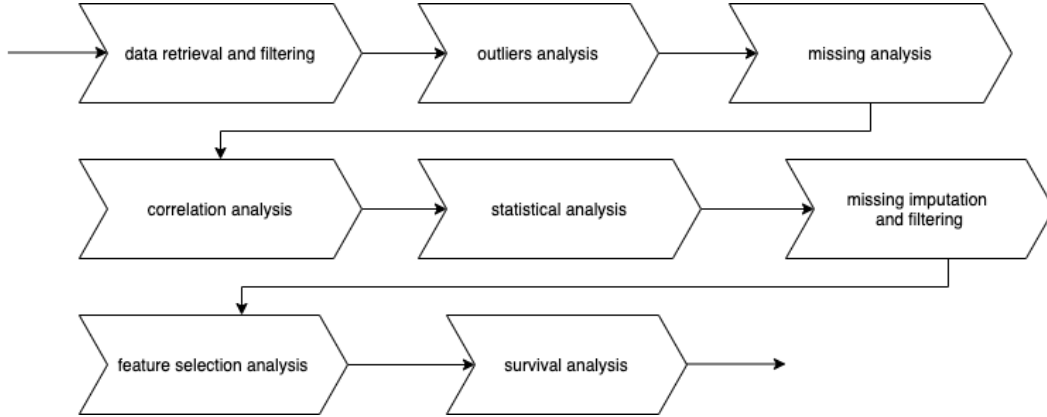


Figure 4.1: Global pipeline

After the pre-processing, we find the statistical analysis and the survival analysis. Given their importance for this work, these two will be discussed later in 5 and 6. Due to their relevance, feature engineering 4.1.6 and error handling 4.1.7 tasks have also been added in this section.

4.1.1 Data Filtering

The initial study population consisted of 2,043 patients to whom a set of filters of interest were applied. In particular, patients were collected if:

- They were part of a previous selection of patients (population restriction). The reason why a previous selection of patients was made was mainly to compare the results of the study with other analyses based on the same patient selection performed in the context of CIBERES-UCI-COVID.
- They belong to the first wave within February 1 and July 31, 2020.
- They had outcome and sex with value as this implies minimal data availability.
- They had been positively diagnosed by COVID-19.
- They required mechanical ventilation on the same day of admission to the ICU. The interest is due to the absence of published studies focused on this population.
- They continued with IMV at least until the third day of ICU. This filter is used for the same reason as the previous one.

This process is shown in Figure 4.2, as input and output of each filter, the size of the dataset can be observed in terms of the number of patients.

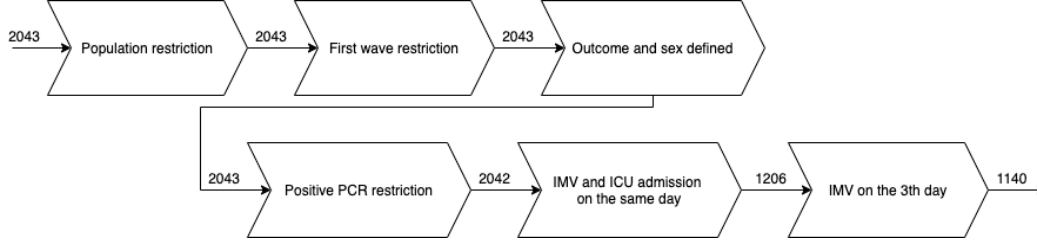


Figure 4.2: Data filtering pipeline

4.1.2 Outliers Analysis

Before performing any statistical analysis, outliers have to be removed. To identify outliers, medical expertise is of capital importance, as they can define the data ranges that can be considered as feasible. There are variables that must be carefully supervised due to their clinical importance. If outliers are located in large numbers, the project coordination must consider the possibility of contacting the hospital in question to understand and fix the source of outliers. For this reason, an analysis of outliers by variable and by hospital is carried out. Initially, extremely high values were found for lactate and platelets. This was due to an error in the unit conversion.

4.1.3 Correlation Analysis

To interpret the relationships that may exist between features, a correlation analysis is performed. Understanding these relationships is useful in order to avoid multicollinearity and redundancy issues that may decrease the performance of machine learning models. According to the types of variables, their correlation can be computed using:

- Pearson for continuous-continuous relations: measure statistical relationship between two continuous variables [21].
- Correlation ratio categorical-categorical: measure of the relationship between the statistical dispersion within individual categories and the dispersion across the whole population or sample [22].
- Cramer's V categorical-continuous and viceversa: measure of association based on Pearson's chi-squared statistic between two nominal variables, giving a value between 0 and +1 (inclusive) [23].

Correlations are computed for all pairs of features in a dataframe and reported to the medical team to decide which variables have higher clinical relevance. Missing values need to be treated now, before moving on since statistical methods cannot handle them. Two strategies to deal with missing data have been implemented. The default strategy is to remove only the missing data taking into account the columns of the current pair. Alternatively we remove all data rows containing missing data. It is preferable to only delete data in one of the two columns that are being treated since otherwise we would be losing large amounts of data.

The current implementation is based on the Dython [24] library with the difference in the treatment of missing values and the possibility to filter correlations by a threshold. The main issue we observed when using Dython was that they just allow to remove missing values dropping entire rows, columns or replacing them. As it is mentioned previously, this causes a significant data loss. For this reason we decided to implement our own library with the same methods as Dython but restricting the missing values elimination only to the pair of variables whose correlation we want to calculate.

Once missing values are ignored in some way or another, correlations are computed and displayed as a heatmap with the value of the correlation coefficient in each cell. In particular, correlations have been calculated for groups of variables: variables related to duration and dates (10.13), demographic data (10.14), comorbidities (10.15), previous medication (10.16), laboratory variables (ICU) (10.17), laboratory variables (3rd ICU) (10.26), mechanical ventilation variables (ICU) (10.35), mechanical ventilation variables (3rd ICU) (10.36), treatments (10.37), complications (10.38) and outcomes (10.39).

Previous plots show in nude and orangish colors positive strong correlations and in light blue negative strong correlations. Correlations can be taken into account if they exceed 0.3 in absolute value. However, only those with high correlation coefficients (> 0.6) will be discussed in this document.

In reference to the length of stay, time in hospital and time in ICU are highly correlated each other as it was expected (10.13). Normally, those patients who spend long periods in the hospital also spent long periods in the ICU. Regarding demographic data, male gender is correlated with female and vice versa (10.14). In Figure 10.15, comorbidities are shown, and no correlations are observed. The situation is similar for previous medication, only the flu, streptococcus vaccine, and *both_vaccine* are correlated. This may be caused by the fact that people who have received one vaccine could also have received the other and vice versa. Also two types of previous respiratory support are lightly correlated (0.54).

For the laboratory and ventilation variables, several correlations have been

found, these have been marked with rectangles and a zoomed in capture has been added in the Appendix. Eight blocks of correlations have been identified for both events numbered C1-C8. Block C1 (ICU) (10.18) shows us correlations between the use of vasopressors and the hemodynamic sofa and sofa scores 0.68 and 0.96 respectively. These correlations are expected since the use of vasopressors is a variable that is part of the calculation of both scores. In Block C2 (ICU) (10.19) we can see the correlation of paCO_2 with ventilatory ratio *modified*. This is expected since the ventilatory ratio *modified* is calculated from paCO_2 , regulated respiratory rate (FR), tidal volume and ideal body weight. Figure C3 and C5 (ICU) (10.20, 10.22) show a fairly high correlation between urea and creatinine. This may be due to increases in both variables at the same time, these increases may indicate the appearance of diseases that affect the liver or kidneys such as hypertension, kidney failure or cirrhosis. Figure C4 and C6 (ICU) (10.21, 10.23) shows several correlations between ferritin, AST, ALT, and troponin-T and lactate respectively. For the first case, high values for ferritin, AST and ALT may indicate that there are patients with liver disease. Lactate and troponin-T are different markers, the first is of tissue perfusion and the other of myocardial damage. They may be correlated in patients who are very critically ill. In Figure C7 (ICU) (10.24) we find the same correlations described by C1 and C2. Finally, Figure C8 (ICU) (10.25) shows several pairs of correlations, all of them expected: hemodynamic SOFA and SOFA (0.71), troponins T and I (1.0), original ventilatory ratio and its modified version (1.0), both enzymes transaminase (0.82) and to a lesser extent APACHE and SOFA scores (0.56) both used to measure the severity of a patient. These same correlations are also displayed for the third day of ICU event. In addition, it is observed that prothrombin time is also related to ferritin in C6 (3r day ICU) (10.32) and in Figure C7 (3r day ICU) (10.33) NT-proBNP presents a high correlation with LDH (0.85) not observed in the event of ICU.

Regarding the mechanical ventilation variables, at a general level in 10.35 and 10.36, it is observed that the negative correlations in the ICU event are accentuated in the event of the third day and new negative correlations also appear. The correlations between PaFi and blood oxygen saturation (from 0.44 to 0.35), and between plateau pressure and driving pressure (0.84 to 0.81) are slightly strengthened. On the other hand, the correlation of RASS and SAS increases (from 0.59 to 0.79). This change between correlations from one event to another may be motivated by the fact that the data from the third day act as a measure of the patient's evolution. Thus, the improvement or deterioration of the patients' condition will be reflected in the variables on the third day of ICU. Taking into account that COVID-19 is a respiratory disease,

it is expected that these changes are observable for mechanical ventilation variables and lose strength or are not observable at all for other variables such as laboratory variables.

With regard to the rest of the variables, no correlations are observed between treatments (10.37) and only one for complications (10.38), DIC (disseminated intravascular coagulation) and coagulation disorder have a positive correlation of 0.45. This is somehow expected since DIC is a coagulation disorder. Finally, Figure 10.39 shows the correlations between outcome variables and complementary therapies. As expected, there are correlations between the 4 types of outcomes: death, transfer to a social-health center, transfer to another hospital and discharge due to improvement.

4.1.4 Missing Analysis

One of the most common situations when working with real data is the existence of missing data. These missing values arise due to many reasons, such as undefined values, data input errors, irrelevant information, mismatch of variables between databases, etc. In the context of CIBERES-UCI-COVID, where data is introduced by dozens of different data entries (each hospital hiring its own), and where hundreds of medical variables are requested for each sample/patient, missing data is frequent and must be addressed thoroughly. Not handling missing data properly can have a negative impact on performance of machine learning models. As the authors of [25] point out, missing values can reduce statistical power and representativeness of samples, introduce bias and reducing drastically the quality of the study.

In view of the considerations above, CIBERES-UCI-COVID have developed a set of techniques that analyses missing data in depth in order to understand its nature and address the issues mentioned before.

Global analysis of missing values In the first place, we analyze missing values from a global perspective using missing maps. The generation of missing maps offers an overview of the amount of existing missing values and their location (both features and samples). Visualising patterns of missing data can help understand the types of missings we are dealing with. In the 10.12 we can see for each variable (x-axis) and patient (y-axis) the variables that are missing (in white). The more white horizontal stripes a column has, the greater the number of missing values for that variable. According to that the variables that have more missing values are:

- Mechanical ventilation variables: SAS (Sedation Agitation Scale) (ICU

and 3th day), compliance (ICU and 3th day), driving pressure (ICU and 3th day), plateau pressure (ICU and 3th day).

- Laboratory variables: NTproBNP (ICU and 3th day), il6 (ICU, 3th day), troponinI (ICU and 3th day), troponinT (ICU and 3th day), ferritin (ICU and 3th day), procalcitonin (3th day).
- Previous medication variables: streptococcus vaccine.

All the previous variables has a missing values percentage equal or above 40%.

Missing values correlations In addition, we generate a correlation matrix heatmap that shows which missing values are correlated with other missing values. Analysing correlations between missing data can help us discover relationships between them (*e.g.*, if the absence of a feature depends in turn on the absence of another). In the Figure 10.1, the regions with strongest correlations are captured using a red frame and a notation above. It has been considered as high correlations those that are equal or greater than 0.5. According to this, a total of 10 regions have been considered.

In 10.2, correlations can be observed between various treatments and previous medication. Specifically, we can realize that when the use of ECMO is missing, so is the use of neuromuscular blockers at least 90% of the time. The same situation is observed between streptococcus and flu vaccine and the statins and heparine. In Figure 10.3, the use of ECMO and neuromuscular blockers appears again showing high correlations to corticosteroid and antibiotic treatments. In turn, antibiotics seem to have a slight correlation with the use of vasopressors and ECMO but only for the third day of ICU (10.9). Neuromuscular blockers also show a correlation with the use of non-invasive ventilation methods (ICU) in 10.7, and with the use of vasopressors and ECMO (third day of ICU) in 10.8. Most expected, in case that vasopressors and ECMO have not been required in ICU they are not required for the third day either (10.10).

According to comorbidities, in 10.4, we can see that there is a subset of them for which if one is missing all the rest are missing as well. This may be due to the fact that during the consultation with the patient, the doctor may decide not to record some comorbidities if the subject does not present certain symptoms. We observe the same behavior for the different types of previous respiratory support depicted in Figure 10.5. In the case of laboratory variables, we found high correlations for lymphocytes and platelets (Figure 10.6) for the ICU event. The correlations for the third day of ICU are accentuated for those variables commonly present in the laboratory tests (10.11).

It may be possible that in some cases these missing data are the result of a dependency relationship, so that if a subject is not measured or prescribed A then neither is measured / prescribed B and therefore there is a correlation between A and B.

Analysis of missing values per variable In more detail, the system analyses the percentage of missing values for each feature over the total samples. It has been observed that the percentage of missing values per variable has decreased as migrations have been carried out. This is because CIBERES-UCI-COVID has given preference to completing existing patients on the REDCap platform rather than adding new patients. The variables with the highest number of missing values are those listed in 4.1.4.

To have an overview of how completed the variables are, apart from listing the percentage of missings individually, we can group the variables according to predefined intervals. We have established intervals between pairs 0, 25, 50, 75, 100%. The number of variables found in each interval is shown below.

- Variables with num. missing values $< 25\%$: 143.
- Variables with num. missing values ≥ 25 and $< 50\%$: 16.
- Variables with num. missing values ≥ 50 and $< 75\%$: 11.
- Variables with num. missing values $\geq 75\%$: 8.

The vast majority of variables (143 out of 179) have a low percentatge of missing values, and there are only 8 variables which are not completed at least 75%.

Analysis of completeness of samples Something relevant for clinical studies is the amount of *useful* patients that are available. This allows to define the sample size of the study and assess whether it is enough to continue, it is necessary to request more samples or if the study must be reconsidered. The degree of completeness of a sample is defined as the percentage of the features that are not missing over the total number of features. This metric is also observed in the right margin of Figure 10.12. As an acceptable limit, patients completed at least 80% have been taken into account. These correspond to a total of 1,019 patients, 121 patients less than at the beginning of the study.

Analysis of missing values grouped by variable In cases where the data is provided by different sources, it may be interesting to analyse the missing data according to their source. In our case, it is useful to analyse these values according to the hospital from which the patients come. In this way, we generate a report where in each row that represents a hospital, the missing data for each variable is detailed. Thus, we know who to turn to when it is necessary to request new data or report an error.

Filtering methods prior to Missing Imputation Before carrying out any imputation technique, we must verify that the current number of missings is feasible as well as establish to what extent the samples must be completed. [26] states that if the proportion of missings is too large it should be considered to simply report the results of the full case analysis and then clearly discuss the interpretive limitations resulting from the trial results. In this regard, we provide methods to filter observations according to the level of completeness and to filter features according to the percentage of missing values.

Patients completed below 80% and those variables with more than 40% of missing values are excluded. At the end, the missing values related to complications, comorbidities, treatments, previous medication and immunodeficiency, streptococcus + gripe vaccines, septic shock for ICU and third day of ICU are filled up. For these cases, we assume that if the value is not known then these treatments have not been administered. For comorbidities we assume they have not taken place.

4.1.5 Missing Imputation

Multivariable imputation by chained equations (MICE) [27] is a particular multiple imputation technique that has become in one of the main methods to address missing data. In this process, missing values are imputed based on the observed values for a given individual and the relations observed in the data for other subjects, assuming the observed variables are included in the imputation model. Intuitively, multiple imputation can also be seen as a process that involves filling in the missing values multiple times, creating multiple *complete* datasets.

It's many advantages have made it more popular than other methods. In the first place, MICE allows to handle variables of different types, and patterns of varying complexities. In addition, it is very flexible allowing a broad range of settings. Unlike other methods like single imputation (e.g mean or mode imputation), MICE allows to take into account the statistical uncertainty in those imputations. Maximum likelihood methods are also a viable approach

but it is not always applicable since only works for certain models and finding implementations for any system different to SPSS is quite difficult. In the case of complete case analysis, it can only be applied under very specific circumstances (e.g. when there is less than 5% missingness and the missingness is totally random and does not depend on observed or unobserved values).

MICE operates under the assumption that given the variables used in the imputation procedure, the missing data are Missing At Random (MAR), which means that the probability that a value is missing depends only on observed values and not on unobserved values. Implementing MICE when data are not MAR could result in biased estimates. In [27], they state that all relationships that are going to be investigated in the analysis need to be included in the imputation model. However, including additional variables that may be not used in further analysis can improve the imputations. This is because they can reduce bias and make MAR assumption, which is almost always impossible to test, more probable.

In short, the chained equation process could be divided in seven steps. In a first step, a single imputation is applied for each missing data. Subsequently, the imputed variables return to their original missing value, the imputed value is saved. For each variable with missing data, a regression is carried out where the rest of the variables with value represent the independent variables and the variable with missing values is the dependent variable. The missing values is thus replaced by the result of the regression. When the variable with initially missing values is later used as the independent variable in regression models for other variables, both the observed and imputed values are used. This process is repeated a predefined number of cycles. After one cycle, all the missing values have been replaced once using regression results. The idea of running several cycles is to stabilize the variables involved in the imputation prediction (regression coefficients). Usually the number of cycles is set to 10. After executing all the cycles, the algorithm is repeated several times to generate multiple datasets. In practice, 5-10 datasets are usually generated. This parameter is also chosen by the user.

The implementation that has been used for this work is from the *fancyimpute* [28] library. To make sure imputations remain within the normal ranges and thus not generate outliers, we pass as an argument a dictionary with the minimum and maximum value of each feature. In Figures 10.49, 10.50, 10.51 the number to be imputed for each variable and the normal ranges considered for each variable is shown sorted in descending order by those that have higher number of missing values. For the cases in which normal ranges are undefined we just set the minimum value to 0. Laboratory and ventilation variables are observed indistinctly, being the highest number of MVs to be imputed 399

(APACHE score for 3rd day) and 1 being the minimum. Furthermore, usually we find the variable for both ICU events. See also that variables days between symptoms appear until ICU admission and time in ICU have been added to the imputation process. This has been the case since the number of patients with MVs for these variables are 1 and 9 respectively, which represents less than 9% of the population. For higher percentages, it may be considered to ignore these patients. Note that the reason for which these patients do not have values for them is because the dates needed to compute those features are missing, symptoms start date and ICU leave date.

The imputations have been validated by comparing feature central tendency measures before and after the imputation for each. Additionally, for numerical features the density plots are compared before and after the imputation. A density plot is nothing else than a representation of the distribution of a continuous variable. The idea is to check that the current distribution is similar to the previous one with missing data without any sudden changes.

The Appendix shows the density plots for each continuous feature that presented missing values. In general we can see that the distributions remain intact before and after the imputation for most variables: creatinine, glucose, HCO₃, heart rate, days between symptoms appear and hospital admission, leucocytes, lymphocytes, paCO₂, paFi, positive pressure at the end of the expiration date (PEEP), platelets, temperature, time in ICU, oxygen saturation and ventilatory ratio.

The distributions remain with small changes in comparison to the original ones for breathing rate, CRP, D-dimer, regulated respiratory rate (FR), Richmond Agitation Sedation Scale (RASS), SOFA score, hemodynamic SOFA, total bilirubin, urea, alanine transaminas (ALT), lactate dehydrogenas (LDH) and ventilatory ratio *modified*.

For APACHE score, lactate, procalcitonin (ICU), aspartate transaminas (AST) and prothrombin time, we find that the distribution with imputed values presents a greater number of subjects with values that are in the median of the variables.

4.1.6 Feature Engineering

After the collection and transformation of REDCap data, we compute derived medical variables and the enrichment of other ones through domain knowledge. It is the case of scoring systems such as APACHE or SOFA score used to measure the critical state of a patient. Most derived variables are computed and saved in the database after migration. There are others that are used only in very particular studies, so it is not necessary for them to be saved

permanently. For instance, the deltas of laboratory and ventilatory variables between the first and the third day ICU are computed at execution time.

Some variables are enriched with knowledge from medical experts. This includes variables that can change their value when some conditions are fulfilled. Usually, these are variables that are involved in the calculation of other more complex variables. This is the case for example of the Glasgow Coma Scale (if it equals 15 for some medical event then the rest of the events take this value), average pressure (if the average pressure read from REDCap is a null value then it is computed using the systolic and diastolic pressures) and modified ventilatory ratio that uses in turn a modified version of tidal volume (if it is a null value, the expiratory volume is used in turn).

4.1.7 Error handling

The error handling is basically divided into two tasks: controlling the errors that may occur in our system and those derived from the REDCap data. For the former, as mentioned above, we perform unit and integration testing. This allows us to quickly detect and resolve bugs, refactor and improve the code, reduce complexity and ensure that all code meets quality standards before it is deployed.

In the case of REDCap errors, our task is to inform the centers so that they can be corrected in the platform. Among the errors found we distinguish the following:

- Variables with wrong units: Although a unit is selected in REDCap, the variable is filled using a different one.
- Variables with impossible values: The variable has a value that is very far from the normal range. These errors are often not outliers but rather typos, which makes them very difficult to correct unless reports or information on other platforms are available.
- Variables with inconsistent values: This happens in variables whose value may be incorrect when observed in combination with others, either because they are part of a sequence or that depend on other variables. Numerous consistency errors have been found regarding dates. For example, some event dates are impossible, which means that some events happen before others or do not refer to the name of the event (third day of ICU does not correspond to the third day but to another).

The risk that we encounter these types of errors is high when data is entered manually, so we must be careful when processing it before conducting any analysis.

4.1.8 Feature Analysis

The goal of carrying out feature selection analysis is to be able to determine those variables that have a greater impact predicting the outcome in order to rule out those that have no predictive force. In turn, this process reduces the size of the problem which helps algorithms to work faster and make models easier to interpret.

Additionally, it has also been investigated whether the fact of imputing missing values causes changes in the relevance of the variables. Initially, different classification and regression algorithms were used and later survival forests algorithm were explored, for which an entire chapter has been dedicated 6. The algorithms have been extracted from the *SKlearn* [29] and survival models have been extracted from *pySurvival* [30]. It has been used logistic regression, random forest and recursive feature elimination validated using 10-fold cross validation for the same subset of features twice, once for imputed missing values by MICE and once for missing values replaced by 0. The performance is evaluated in terms of both accuracy and f1-score (the harmonic mean of the f1-scores obtained by each class is compared). On the sidelines, the Information Value (IV), commonly used in marketing, has also been tested.

Before the analysis, continuous variables are standardized and date type variables are ignored throughout the process. Variables that had been eliminated due to a considerable number of missing values and those whose clinical weight was not considered relevant were also ignored. However, the possibility of adding the ones with less clinical weight in subsequent studies or models was considered. Specially, the variables that have been ignored and have passed the filters prior to imputation can be seen in Table 4.1.

Missing imputation and correlation analysis tasks have been carried out with a larger set of features for several reasons. First, for exploration purposes. At the beginning of the study, there was no record of the list of variables to be analyzed and the main objective was to analyze the data itself and carry out a sanity check to see if expected correlations were obtained. Second, and as mentioned above in 4.1.5, running the imputation on a larger set of features helps to reduce bias of the imputations.

The features that have reached the greatest importance in predicting the outcome are shown below for each algorithm.

Topic	Metrics
Patient characteristics	-
Previous Medication	alpha blockers, reninInhibitors, vasodilators, antimineralocorticoids, adrenergic antagonists, AINE, heparine
Comorbidities	smoker
Laboratories (only ICU and 3rd day ICU)	il6, vasopreossor requirement, ECMO requirement, heart rate, temperature, HCO3, prothrombin time, bilirubin, aspartate transaminase (AST), alanine transaminase (ALT), troponin-T, troponin-I, NT-proBNP, APACHE score
Mechanical Ventilation (only ICU and 3rd day of ICU)	Richmond Agitation Sedation Scale (RASS), regulated respiratory rate (FR), Riker sedation agitation scale (SAS), driving preassure, plateau pressure, positive pressure at the end of expiration date (PEEP), breathing rate, ventilatory ratio modified
Hospital Course	-
Complementary Therapies	oxigenotherapy required, tracheostomy required, ECMO required
Treatments	-
Complications	viral pneumonia, heart attack, coagulation disorder, sdra, pancreatitis, skin manifestations, rhabdomyolysis, convulsions complications, cardiomyopathy, meningitis, anemia
Outcomes	-

Table 4.1: List of variables classified by topic ignored before performing feature analysis

Logistic regression The variables that are most important when predicting the death event are time in the ICU, age, respiratory, coronary and renal comorbidities (cardiac arrest, acute kidney failure, pulmonar chronic disease, renal chronic disease), complementary treatments and therapies (corticosteroids and neuromuscular blockers). Regarding the ventilation variables, the ventilatory ratio for both events is the one that receives the greatest importance. when the missing values are replaced, it is observed that the importance of the variables varies but without producing significant changes.

Random forests As can be seen in Figures 10.115 and 10.116, in both cases the variables that receive the greatest importance are time in hospital, age, time in ICU, and a whole list of laboratory variables corresponding to the third day of ICU (paFi, platelets, paCO₂, creatinine, LDH, urea and SOFA score). This last pattern is clearer for the imputed data. As previously discussed, this may be due to the fact that the third day variables constitute an indicator of the evolution of the patient's condition.

Recursive feature elimination The results for this algorithm are very similar to those obtained when the estimator is used directly. This is expected, although recursive feature elimination removes one or more features in each iteration, the chosen estimator is the same.

Information Value (IV) Information value helps to rank variables on the basis of their importance using the weight of evidence (WoE). Low information value of a certain variable indicates that the predictive power of this variable is low and thus it is not able to classify the target variable properly. The weight of evidence measures the predictive power of an independent variable in relation to the dependent variable. It has its roots in credit scoring world and it tells the degree of the separation between subjects than can belong to two different classes.

According to [31], by convention the values of the IV can be interpreted as follows:

- Not useful for prediction: Less than 0.02.
- Weak predictive power: 0.02 to 0.1.
- Medium predictive power: 0.1 to 0.3.
- Strong predictive power: 0.3 to 0.5.

method	missing imputation applied	missing values replaced by zero
logistic regression (L1)	0.90	0.91
random forest	0.91	0.87
recursive feature elimination with random forest	0.88	0.89
recursive feature elimination with logistic regression	0.87	0.83

Table 4.2: Performance in terms of accuracy of each algorithm with missing imputation and without imputation replacing missing values by zero

- Suspicious or too good predictive power: > 0.5 .

Figures 10.121 and 10.122 show the predictive power of each variable. Note that there is no variable that exceeds 0.05, which leads us to consider that IV is not working correctly.

One of the causes that can explain this phenomenon is that there are numerical variables incorrectly binned. WoE uses binned numerical variables and on those bins the log odds ratio is calculated. Too few bins causes harsh aggregation that loses much of the available information from that variables while too many may cause overfitting.

In general terms, we can conclude that the selection of features varies from one method to another, as is logical, but characteristics such as time in ICU, age and certain comorbidities appear at the top of all the methods, ignoring IV. It is also observed that the fact of replacing missing values by zeros does not introduce significant changes in the predictive power of the variables.

To complement this work, Table 4.2 shows the levels of accuracy achieved omitting IV. Again, the replacement of missings also has no effect on the performance of the estimators.

Based on the results of the logistic regression, random forest and recursive feature elimination and expert medical knowledge, those variables that are considered to have the greatest clinical weight have been selected. Table 4.3 shows the final set of features.

Topic	Metrics
Patient characteristics	age
Previous Medication	flu vaccine, statins, interferon beta, ARB, ACE-inhibitors
Comorbidities	hiv, asthma, hypertension, heart chronic disease, pulmonar chronic disease, renal chronic disease,
Laboratories (only ICU and 3rd day ICU)	platelets, creatinine, LDH, D-dimer ICU, lymphocytes, septic_shock, lactate, urea, oxygen saturation. hemodynamic SOFA score
Mechanical Ventilation (only ICU and 3rd day ICU)	previous respiratory support, simple face mask (Hudson), ventilatory ratio, paFi, paCO2
Hospital Course	time in hospital, time in ICU, time IMV
Complementary Therapies	neuromuscular blockers requirement
Treatments	antibiotic, corticosteroid, tocilizumab
Complications	cardiac arrest, DIC, pulmonary embolism, ictus, organizing pneumonia, bacteremia, lung infectious complications, hypertension, acute kidney failure
Outcomes	alive28Days

Table 4.3: Set of selected variables after feature selection using conventional methods

Chapter 5

Statistical Analysis

A statistical analysis has been performed prior to survival analysis in order to interpret the data and discover patterns and trends. Specially, it has consisted in the elaboration and interpretation of a total of seven tables in which we examined: Characteristics of the patients according to age, gender, previous medication and appearance of symptoms (10.40), comorbidities and symptoms prior hospitalization (10.41, 10.42), characteristics on admission in ICU first and third day (10.43, 10.44), characteristics at the beginning and end of mechanical ventilation (10.45, 10.46), treatments and complications during hospital stay (10.47), and outcome variables and complementary therapies (10.48).

For each table, categorical variables are presented as frequency/percentage of a group from which they were derived, and for continuous variables the median [interquartile range (IQR)] is shown. Categorical variables were compared with the use of Chi-square test or Fisher's exact test, while continuous variables were compared with the Student's t test or Mann-Whitney U test. Missing values for each feature are ignored.

During the study period, 1,140 critically ill patients with COVID-19 were admitted to the ICUs under the conditions aforementioned. The median age is 65 years (IQR 56–71), and 816 (71.58%) are male. Of these patients, 30.42% received statins, 17.63% ACE inhibitors, 16.58% ARB and 14.82% diuretics among common previous medication. The median duration from symptom onset to hospitalization is 7 days (IQR 5–10). Compared with survivors, non-survivors were more likely to be older [69 (IQR 62.5–74) vs. 61 (IQR 53–69), $P < 0.05$] and male (75.62% vs. 69.01%, $P < 0.05$). The length of ICU stay was significantly longer for survivors (10.40).

Hypertension (49.78%) appears as the most common comorbidity, followed by obesity (33.66%) and diabetes (22.83%). Important differences between

survivors and non-survivors are observed for chronic kidney disease [3.73% vs. 7.69% $P < 0.05$], heart chronic disease [8.61% vs. 17.87% $P < 0.05$] and pulmonary chronic disease [7.75% vs. 16.29% $P < 0.05$] (10.41). Among the most common symptoms we find fever (88.8%), dry cough (67.41%), shortness of breath (71.44%), fatigue (33.84%) and muscle pain (25.88%). In general, it is observed that the percentage of patients showing symptoms is higher in survivors (10.42).

Laboratory findings suggest that leucocytes count, CRP, LDH, d-dimer, NT-proBNP, urea are higher in non-survivors suggesting more severe systemic inflammation, cell injury, coagulopathy, risk of cardiac failure and uremia. As suggested by higher APACHE score, non-survivors were more severely ill (10.43). These differences are maintained on the third day of ICU. The severity scores APACHE and SOFA score are now more distant, being the values for survivors equal or less than the first day of ICU. The use of vasopressors is accentuated for non-survivors on the third day, however, for survivors, IL-6 levels fall, suggesting an improvement in the clinical prognosis (10.44).

Regarding the ventilation variables on the first day of mechanical ventilation, no differentiating facts were observed between survivors and non-survivors. Note only that $paFi$ levels are slightly higher for survivors [118 vs. 112.5 $P < 0.05$] (10.45).

At the end of the treatment, a general deterioration trend is seen for non-survivors. Paused respiratory rate, driving pressure, plateau pressure, ventilatory ratio modified and not modified and $paCO_2$ increase while compliance, $paFi$, oxygen saturation decrease with respect to the first day of ICU. The reversed trend is observed for survivors. This also reflected into low or no sedation levels for survivors and very deep for non-survivors [RASS 0 (IQR -1-0) vs. -5 (IQR -5-4), SAS 4 (IQR 4-4) vs. 2 (IQR 1-2)]. This general deterioration is also reinforced by the differences between SOFA [3 (IQR 2-4) vs. 8.5 (IQR 6-11)], hemodynamic SOFA [0 (IQR 0-0) vs. 4 (IQR 1-4)] and APACHE score [9 (IQR 7-12) vs. 18 (IQR 14-22)] that increase considerably between both groups (10.46).

Regarding complications, sepsis (86.12%), anemia (72.81%), hyperglycemia (72.63%), chronic kidney disease (43.77%), lung infection (42.46%), bacteremia (40.28%) are the most common complications during hospitalization. In response, the most used treatment has been antibiotics (99.12%) followed by corticosteroids (76.46%). Pneumothorax [8.32% vs. 14% $P < 0.05$], cardiac arrest [1.87% vs. 17.83% $P < 0.05$], bacteremia [38.36% vs. 43.31% $P < 0.05$], coagulation disorder [23.99% vs. 31.22% $P < 0.05$], chronic kidney disease [34% vs. 59.14% $P < 0.05$], hemorrhages [7.77% vs. 15.12% $P < 0.05$] occur to a greater extent for non-survivors. No differences are observed in the treatments for any

of the groups (10.47).

Of the 1,140 patients, 443 (38.86%) died, 585 (51.32%) were discharged, 74 (6.49%) were transferred to health centers and 38 (3.33%) were transferred to other hospitals. It is ensured that these transfers are for follow-up purpose, family request and other factors that do not include relapse of the patient. The percentage of non-surviving patients who survive 28 days before dying is only 17.46%, which indicates that the majority of non-survivors do not reach one month of life. Specifically, non-survivors, despite receiving mechanical ventilation for the same number of days as survivors, spend less time admitted to the ICU [22% vs. 16% $P < 0.05$] and in the hospital in general [38% vs. 19% $P < 0.05$]. Non-survivors also tend to need reintubation more often than survivors [89.41% vs. 95.31% $P > 0.05$]. According to complementary therapies, the number of times a patient has been placed in the prone position, has required recruitment maneuvers and neuromuscular blockers is higher for non-survivors (10.48).

The results obtained do not differ from other studies done in other countries such as France, Italy, China and those already existing in Spain. The current analysis comprises 1,140 critically ill patients with COVID-19 whose median age is 65 years [(IQR) 56-71] and 71.58% are male. The studies carried out during the first wave by [3], [5], [4], and [32] show a similar mean age [69.4 (IQR 18-102)], [67.5 (IQR 10-94)], [66 (IQR 20-100)], [65 (IQR 56-73)]. Regarding gender, there is a general trend that the percentage of men is always higher than women's 59% [3], 63.63% [5], 59% [4], and 65.1% [32].

It should also be noted that the set of comorbidities remains fairly stable among the different countries. In first position we find hypertension as the most common comorbidity 50.9% [3], 34.09% [5], 52.4% [4], and 42% [32]. Diabetes is the second most common for 39.7% [3], 25.3% [4], and 18.8% [32]. For [5] heart disease 25% was reported. Alternatively, [5], [4], [32] report cardiac diseases and the last two report chronic kidney disease.

Also, the main symptoms do not change between countries. The most common symptom is fever 84.2% [3], 78.1% [4], and 85.9% [32] and 90.90% [5] followed by cough 73.5% [3], 68.4% [4], 75% [32] and 90.90% [5]. Third, shortness breath is observed for 57.6% [3], 66.1% [4], and 60.7% [32] and 22.74% [5].

The main differences between these studies, leaving aside the number of samples used, lies in the laboratory findings. [3] states high values of ferritin (73.5%), lactate dehydrogenase (73.9%), and D-dimer (63.8%) as well as lymphopenia (52.8%) were frequent. [5] also reports a low number of lymphocytes, leucocytes and elevated LDH values in addition to thrombocytopenia. [32] observe patterns similar to ours, laboratory tests show that leucocytes,

CRP, LDH, hsc-TnI, and D-dimer are higher in non-survivors. After 14 days in ICU, paFi and lymphocyte count steadily improve in survivors and remain low in non-survivors. In comparison, hs-CRP and LDH levels significantly decrease in survivors but remain higher in non-survivors. D-dimer and hsc-TnI levels are relatively stable, but significantly higher in non-survivors than survivors [32].

Regarding mechanical ventilation, [4] reports that 28.1% of patients used IMV and 21.4% required vasopressors. However, the percentages shown in [32] are higher, it is stated that 52.9% patients required NIMV, 41.9% IMV and 40.4% required vasopressors. These values are considerably lower than those shown in this study. In our case, there is a higher tendency for patients to receive IMV instead of NIMV for both events and require vasopressors (66.52% for ICU and 68.76% for third day of ICU). As expected, the most common treatments across the studies are antibiotics and antivirals.

With regard to mortality, the percentages change across countries taking into account patients only from the first wave. In our study, the percentage of non-survivors corresponds to 38.85% while [4] (carried out in the north of France) reports 11% and [3] (carried out in Spain) reports 21%, being those older than 80 years those who die in higher proportion (46.0%) followed by those over 70-79 (26.9 %).

In conclusion, we can confirm that the most common comorbidities, symptoms and treatments remain the same between countries but there are differences in the clinical biomarkers, the durations of IMV, NIMV, and the resulting mortality rates.

Chapter 6

Survival Analysis

Survival analysis can be defined as the methodologies used to explore the time it takes for an event to take place. It involves the consideration of the following variables:

- Time of origin: time at which patient follow-up begins.
- Time of event: time at which the event occurs.
- Time to event: difference between time of event and time of origin, also called failure time, survival time, or event time.
- Event: e.g., death, failure, etc.

Unlike regression, in which we have a continuous outcome, and classification in which we have a set of discrete labels, in survival analysis the outcome is associated with both the event and the time value. The fact that the time to event does not follow a normal distribution and that the time to event may be unclear for some of the subjects, makes conventional methods unsuitable for this type of data. The survival time response may be incompletely determined for some subjects, those subjects are called censored subjects. Essentially, censoring is present when we do not know the exact event time for that subject. As it is pointed out in [33], there are three main reasons why this happens: individual does not experience the event when the study is over (*i.e.*, survivors), he/she is lost to follow-up during the study period. or withdrawn from the study.

According to that, there are three types of censoring:

- Right censoring: the most common censoring, the survival time is incomplete at the right side of the follow-up. In other words, it is known

that patients survived up to the time they withdrew, but we do not know the exact survival time. This happens for those subjects that did not experience the event by the end of the study or that withdrew before the study ended. The true patient survival time in right censoring will be always equal or greater than the observed survival time.

- Left censoring: the subject experiences the event before the study starts but we do not know when. In that case, patient survival time is less than or equal to the observed survival time.
- Interval censoring: the subject experiences the event between two moments of time inside the study but we do not know exactly when the interval starts or ends.

These phenomena are exemplified in Figure 6.1 in which we can observe two patients left-censored (in green) and two right-censored (in blue).

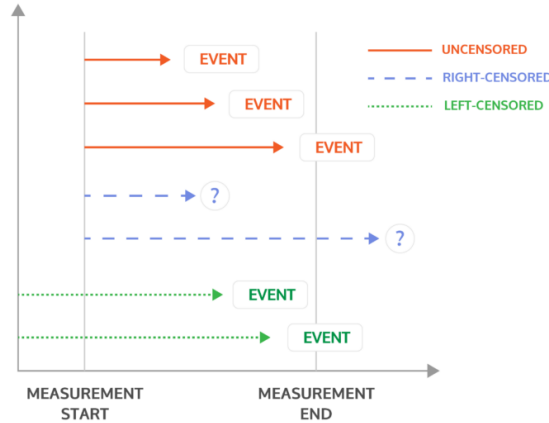


Figure 6.1: Types of censoring [34]

Survival data is generally modelled in terms of two related functions: survival and hazard. In [33], the authors describe that the survival probability, also called the survivor function $S(t)$, as the probability that an individual survives from the time origin (e.g. diagnosis of cancer) to a specified future time t .

$$S(t) = P(T > t)$$

The hazard is usually denoted by $h(t)$ or $l(t)$ and is the probability that an individual who is under observation at a time t has an event at that time. In other words, it represents the instantaneous event rate or potential risk for an individual who has already survived to time t to experience the event in t .

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

If we pay attention to the hazards formula, we can observe that if either $S(t)$ or $h(t)$ is known, the other can be determined automatically. However, there is no simple method to estimate $h(t)$ directly. Instead, $h(t)$ is estimated using other measures such as the cumulative hazard $H(t)$. This can be defined as the integral of the hazard, or the area under the hazard function between times 0 and t . According to [33], it could be interpreted as the cumulative force of mortality, or the number of events that would be expected for each individual by time t if the event were a repeatable process.

In summary, survival probabilities allow us to know the probability for an observation to survive longer than a certain time t , which allow in turn to describe the survival experience of a study cohort. In contrast to the survivor function, the hazard function provides insight into the conditional failure rates and provides a vehicle for specifying a survival model. Actually both have opposite meanings, survivor function focuses on not having the event while hazard function focuses on the event occurring.

6.0.1 Models

6.0.1.1 Kaplan-Meier Estimator

The Kaplan-Meier estimate is the simplest way of computing the survival probabilities over time from observed survival times [35], for both censored and uncensored observations. The survival curve is defined as the probability of surviving in a given length of time while considering time in many small intervals [36].

To properly use the Kaplan-Meier estimator three assumptions must be met. In the first place, it is assumed that at any time patients who are censored have the same survival prospects as those who continue to be followed. This assumption cannot easily be tested and when it is, it may not be true for all censored patients. In practice, if the percentage for those censored in which the first assumption is not fulfilled is low, this violation will induce only a very limited bias in the survival probability and thus a assumable bias. In [37] they use Kaplan-Meier estimator to examine all-cause mortality in patients starting renal replacement therapy (RRT) and in the context of patient censorship,

patients were censored at the time of recovery of renal function. Since these patients may be healthier, they may have better survival prospects than the patients still in the study. However, as only 1.3% of the study participants recovered renal function, this unfulfillment will induce a very small bias in the survival probability on RRT estimates.

Another important assumption for censoring is that the survival probabilities should be the same for patients who were recruited early and patients who were recruited late during the study period. This assumption can be tested by comparing the survival curves for patients who were recruited early and those who were recruited late. If survival curves are similar then it can be concluded that the assumption is fulfilled.

Thirdly, it is assumed that the event happens at the time specified. This creates a problem in some conditions when the event would be detected at a regular examination [35]. For this reason, survival can be better estimated if the time intervals for which data are available are shorter.

As it is shown in the next formula, for each time interval, survival probability is calculated as the number of subjects surviving divided by the number of patients at risk.

$$S_t = \frac{\text{Number of subjects living at start} - \text{Number of subjects died}}{\text{Number of subjects living at the start}}$$

Later, to compare survival curves between groups the log-rank test can be used. This statistic allows to test whether those curves are statistically different or not. However, it does not allow to test the effect of the other independent variables reason for which is considered as an example of univariate analysis. They describe the survival according to one factor under investigation, but ignore the impact of any others. Additionally, Kaplan-Meier curves and logrank tests are useful only when the predictor variable is categorical but they do not work easily for quantitative predictors such as gene expression, weight, or age.

6.0.1.2 Cox's proportional hazard regression (CoxPH)

An alternative method to Kaplan-Meier estimator is the Cox proportional hazards regression analysis, developed in 1972, this model is still one of the most important methods used for modelling survival analysis data. Unlike the previous model, Cox regression model works for both quantitative predictor variables and for categorical variables. Furthermore, it extends survival analysis methods to assess simultaneously the effect of several risk factors on survival time.

The purpose of this model is to evaluate simultaneously the effect of several factors on survival. In other words, it allows us to examine how specified factors influence the rate of a particular event happening (e.g., infection, death) at a particular point in time. This rate is commonly referred as the hazard rate and predictor variables called covariates.

The Cox model is expressed by the hazard function denoted by $h(t)$. It can be estimated as follow:

$$h(t) = h_0(t) * \exp(b_1x_1 + b_2x_2 + \dots + b_px_p)$$

where t represents the survival time, the coefficients (b_1, b_2, \dots, b_p) measure the impact of covariates and the h_0 that is the baseline hazard. It corresponds to the value of the hazard if all the covariates x_i are equal to zero. The quantities $\exp(b_i)$ are called hazard ratios (HR). If the value of the coefficient is greater than zero, or equivalently a hazard ratio greater than one, this indicates that as the value of the i^{th} covariate increases, the event hazard increases and thus the length of survival decreases. Put another way, a hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival. This is why, usually a HR greater than 1 is called a bad prognostic factor.

A key assumption of the Cox model is that the hazard curves for the groups of observations (or patients) should be proportional and cannot cross [38]. Consequently, the Cox model is a proportional-hazards model: the hazard of the event in any group is a constant multiple of the hazard in any other. This assumption implies that, as mentioned above, the hazard curves for the groups should be proportional and cannot cross. If an individual has a risk of death at some initial time that is twice as high as that of another individual, then at all later times the risk of death remains twice as high.

These assumptions should be tested prior to application of Cox regression analysis routinely. One possible way to test it by using the examination of the Kaplan–Meier curves. If there is a crossing of the Kaplan–Meier curves of the two groups or the curve of one arm drops down, while the other plateaus then there is a violation of the assumption [39]. Another possible way is to use the Scaled Schoenfeld residuals, statistical tests and graphical displays that check the proportional hazard assumption [39].

The evaluation of the proportional hazards assumption is essential since its violation raises questions regarding the validity of Cox model results which, if unrecognized, could result in the publication of erroneous scientific findings [40]. Although others argue that checking this assumption is only necessary if your goal is inference or correlation. Lifelines documentation mentions that

if your goal is prediction, checking model assumptions is less important since your goal is to maximize an accuracy metric, and not learn about how the model is making that prediction.

6.0.1.3 Cox’s time varying proportional hazard model

In [41], the authors defend that in those situations in which the proportional hazards assumption of the Cox regression model does not hold, the effect of the covariate is then time-varying. In other words, the fact that the null hypothesis is rejected induces the use of time varying coefficient to describe the data. In order to identify time-varying coefficients, we should test the proportional hazards assumption after fitting a Cox proportional hazard model.

Having covariates that change over time during the follow-up period is a common phenomenon in clinical research. For example, the effect of smoking on cancer risk has been extensively studied. However, the smoking status is ever changing during the follow up period [42]. Such a covariate can be considered as a time-varying covariate.

Time-varying covariates can be classified as either internal or external. The first to carry out this distinction were [43]. They define an external covariate such as the one that is not directly related to the failure mechanism. For instance, the age of an individual in a long-term follow-up study is considered an external covariate. On the other hand, an internal covariate is a value over time generated by the individual under study. An internal covariate is for example the procedural history of a patient and the variables measured in a test result. This classification is helpful in interpreting the regression models and results for time-dependent covariates.

The key assumption of including time-varying covariates is that its effect does not depend on time. Time-variant features should be used when it is hypothesized that the predicted hazard depends significantly on later values of the covariate than the value of the covariate at the baseline. To introduce time-varying covariates, the Cox proportional hazard model is extended. The general mathematical description is:

$$h(t|x) = h_0(t) \exp(\sum_{i=1}^n \beta_i(x_i(t) - \bar{x}_i))$$

Note that now, covariates are described using a parametric time function x_i to indicate that they change over time. One way to indicate the different values that a time-varying covariate may have during the follow-up is to use a step function. The idea of this method is to split the analysis time into several intervals. The most common approach to indicate the beginning and ending of each interval in the dataset is to use two additional columns *start* and *stop*.

This pre-processing step is also known as transforming the dataset into “long” format.

6.0.1.4 NonLinear CoxPH: DeepSurv model

Cox’s proportional hazards model described above is a linear model. This model estimates the log-risk function $b_1x_1 + b_2x_2 + \dots + b_px_p$, also known as risk score as a linear combination of features, *i.e.*, using a linear function. This makes cox models unsuitable for analyzing high-dimensional data, low-sample size data and highly non-linear relationship between covariates. The high-dimensional data often result in, either training infeasible or overfitting of the training dataset [44].

Deep learning-based survival analysis has been highlighted due to its capability to identify nonlinear prognostic factors, higher predictive performance and flexible model design [45]. One of the common approaches to build deep learning models for survival analysis is the adaptation of the Cox proportional hazard assumption. For that, those models incorporate a standard Cox-PH model as an output layer. This approach was originally proposed by [46] in 1995. Although, they focused their study on multilayer perceptrons, further research is paying attention to more advanced architectures such as convolutional neural networks and recurrent neural networks.

In the context of DeepSurv [47], the model consists of a deep feed-forward neural network which predicts the log-risk function. This function is parameterized by the weights of the network. The observed covariates are feed into the model as input features. The hidden layers of the network consist of a fully connected layer of nodes, followed by a dropout layer. Finally, the output layer has one node with linear activation, which estimates the log-risk function in the Cox model.

The promising results shown in [47] demonstrate that DeepSurv achieves competitive performance, compared to standard Cox-PH alone and random survival forests. With further research dedicated to this direction, deep learning survival approaches has the potential to substitute traditional survival analysis methods for medical researchers to study and predict the effects of patient’s covariates on their risk of failure.

6.0.1.5 Random Survival Forest

A random survival forest (RSF) is an assemble of trees method for analysis of right censored time-to-event data. This method estimates cumulative hazard function (CHF) by averaging the Nelson-Aalen cumulative hazard function of

each tree [48]. The idea is based on the principle of conservation of events for survival trees, this means that the sum of the estimated CHF over observed time (both censored and uncensored) equals the total number of deaths in fairly and general conditions. Under this principle, a new outcome variable called ensemble mortality is defined. As mentioned before, ensemble mortality is the expected total number of deaths computed by the forest predicted value for the CHF. The principle of conservation of events is applied to a wide collection of estimators, including the Nelson–Aalen estimator.

The algorithm draws in the first place, B bootstrap samples of the same size from the original data. Note that each bootstrap sample excludes on average 37% of the data, called out-of-bag data (OOB data). Similar to CART, survival trees are binary trees grown by recursive splitting of tree nodes. A good split for a node maximizes survival difference between daughters. The best split for a node is found by searching over all possible x variables and split values c , and choosing that x^* and c^* that maximizes survival difference. By maximizing survival difference, the tree pushes dissimilar cases apart. Eventually, as the number of nodes increase, and dissimilar cases become separated, each node in the tree becomes homogeneous and is populated by cases with similar survival. Previous described impurity measure is known as the log-rank score split-rule.

This process is repeated until a stopping criterion is met. Eventually the survival tree reaches a saturation point when no new daughters can be formed because of the criterion that each node must contain a minimum of prespecified number of deaths. The most extreme nodes in a saturated tree are called terminal nodes.

Once the tree is built, a cumulative hazard function (CHF) is computed for each tree using the Nelson-Aalen estimator.

$$\widehat{H}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}}$$

where $d_{l,h}$ and $Y_{l,h}$ are the number of death and individuals at risk at time point $t_{l,h}$. Thus, all observations within the same node have the same CHF. The ensemble CHF is computed averaging over all CHFs for the B trees. At the end, the prediction error for the ensemble CHF is computed using OOB data.

Clearly, it can be seen that this method is nothing else than an extension of classification and regression trees and random forests (RF) for time-to-event data. One of the main advantages of using random survival forest is flexibility and ease of dealing of high dimensional covariate data. On the other hand, it presents the common drawbacks of random forests including a bias towards inclusion of variables with many split points. This effect leads to a bias in resulting summary estimates such as variable importance [49].

6.0.1.6 Conditional Survival Forest

The Conditional Survival Forest model was developed by [50] in 2017 to improve the Random Survival Forest performance. As just mentioned in the previous model, split variable selection of Random Survival Forests favors splits for covariates with many possible split points. If the split variable selection is biased, other parameter estimates, such as variable importance measures are biased as well. Thus, predictions could also suffer from the underestimation of important variables with few categories. This increases the need for unbiased split variable selection.

An approach to avoid split variable selection bias are Conditional inference forests (CIF) [51]. CIF are known to correct the bias in RSF models by separating the procedure for the best covariate to split on from that of the best split point search for the selected covariate. To determine the optimal split variable an association test is conducted. In particular, a linear rank test is used based on the log-rank transformation (log-rank scores). If the association is found to be significant, the covariate with minimal p-value is selected for splitting. If no significant association is found, no split is conducted.

However, linear rank statistics cannot detect non-linear effects in the independent variables. In [50], the authors present an alternative to use maximally selected rank statistics for the split point selection instead of default linear rank test to reduce split variable selection bias. Roughly, the skeleton of the algorithm in general follows the same procedure as RSFs and uses the CIF two-step procedure to split nodes with the difference that initially the optimal binary split is determined as in standard random forests, but an adjustment for multiple possible splits is performed through the use of maximally selected rank statistics. In contrast to CIF, they use [?] approach to adjust for multiple testing and to decide whether tree growing should be stopped.

6.0.1.7 Extremely Randomized (Extra) Survival Trees

The high variance of decision and regression tree splits of previous work motivate the authors of [52] in 2005 to perform a bias/variance analysis to investigate whether higher randomization levels could improve accuracy with respect to existing ensemble methods.

The Extra Survival Trees model is an extension of the Extremely Randomized trees model, introduced by [52]. They propose a new tree algorithm that selects splits, both attribute and cut-point, total or partially at random independently of the target variable. Unlike other ensemble methods, they use the whole learning sample (rather than a bootstrap replica) to grow the trees.

For a given numerical attribute x , selects its cut-point fully at random using N_{splits} values drawn from a uniform distribution over the interval $[\min(x), \max(x)]$. At each tree node, this is combined with a random choice of a certain number of attributes among which the best one is determined. In the extreme case, the method randomly picks a single attribute and cut-point at each node, and hence builds totally randomized trees. Although they also propose a way to select random splits for categorical attributes, the authors focus mainly on the study of this randomization idea in the context of numerical attributes only.

6.0.1.8 Linear SVM

Survival Support Vector Machines are an extension of the standard Support Vector Machine applied to right-censored time-to-event data. Its main advantage is that it can account for complex, non-linear relationships between features and survival via the so-called kernel trick. A kernel function implicitly maps the input features into high-dimensional feature spaces where survival can be described by a hyperplane.

Initially, survival analysis in the context of Support Vector Machines was described as a ranking problem. [53] developed the Rank Support Vector Machines (RankSVMs) in which the model learnt to assign samples with shorter survival times a lower rank by considering all possible pairs of samples in the training data. The idea behind formulating the survival problem as a ranking problem is that in some applications, like clinical applications, one is only interested in defining risks groups, and not the prediction of the survival time.

Few years later, [54] design a straightforward algorithm to efficiently use the primal formulation, by computing a convex quadratic loss function. In that way, Newton optimization can be used it to minimize the loss function. In addition, they extended the same optimization for non-linear models.

$$\begin{aligned} \min L_p &= \frac{1}{2} \|w\|^2 - \sum_{i=1}^l a_i y_i (x_i^T w + b) + \sum_{i=1}^l a_i \\ \text{s.t } \forall i \ a_i &\geq 0 \end{aligned}$$

6.0.1.9 Survival Analysis in the presence of competing events

A competing event is one that precludes the occurrence of the event of interest. They compete with each other to deliver the event of interest, and the occurrence of one type of event will prevent the occurrence of the others.

For example, in a survival study where primary outcome is death by cardiovascular causes, death by noncardiovascular causes is a competing risk. Regardless of how long the duration of follow-up is extended, a subject will

not be observed to die of cardiovascular causes once he or she has died for noncardiovascular causes [55].

Conventional statistical methods for the analysis of survival data assume that competing risks are absent. In fact, if a patient experiences a competing event, standard survival analysis methods treat that patient as censored for the outcome of interest. The event time is unobserved for censored subjects, hence the statistical analysis will proceed without knowledge of the event time for those subjects. All that is known is that their event time occurred after the time at which they were censored.

Conventional statistical methods for the analysis of survival data make the important assumption of independent or noninformative censoring. This means that, at any given time, subjects who remain in follow-up have the same future risk of the event occurring as subjects who are no longer followed either by censoring or by dropping out of the study [55].

Censoring subjects which have experienced an competing event may be problematic. In the first place, this may violate the assumption of noninformative censoring. Those subjects that are still alive may not be able to represent those subjects that have been censored. In addition, censoring subjects which have experienced an competing event may lead to interpret incorrect event probabilities (relevant in many practical applications.) since they were not censored in the original environment.

As a result, conventional methods such as Kaplan-Meier estimator and Cox proportional hazards model lead to biased results. In particular, in [56] they observed that Kaplan-Meier curves overestimate the incidence of the outcome over time and Cox models inflate the relative differences between groups, resulting in biased hazard ratios in the presence of the competing events.

Cumulative Incidence Function (CIF) One of the most popular alternative approaches to analyze competing event data is called the Cumulative Incidence Function (CIF). This method estimates the incidence of each of the different types of competing risks. The incidence for each competing event is measured in terms of its marginal probability. Marginal probability is defined as the probability of subjects who actually developed the event of interest, regardless of whether they were censored or failed from other competing events.

$$CIF_c(t_f) = \sum_{f'=1}^f \hat{I}_c(t_f) = \sum_{f'=1}^f \hat{S}(t_{f'-1}) \times \hat{h}_c(t_{f'})$$

where $\hat{S}(t_{f'-1})$ denotes the estimate of overall probability of surviving at previous time $t_{f'-1}$. Overall survival is taken into consideration because we

need to ensure that a subject must have survived all other competing events in order to fail from event type c at time t_f .

$\hat{h}_c(t_f) = \frac{m_{cf}}{n_f}$ represents the estimate of hazard at ordered failure time t_f for event-type of interest c also called cause-specific hazard. m_{cf} denotes the number of events for risk c at time t_f and n_f is the number of subjects at that time.

$\hat{I}_c(t_f)$ denotes the probability of failing from the event type c at time t_f represented by the product of surviving the previous time periods and the cause specific hazard at time t_f . At the end, the CIF for event type c at time t_f is then the cumulative sum up of all the incidence probabilities of event type c of all previous time until t_f .

Cause-specific hazard model Cause-specific hazard models are used to separately estimate the failure rate for each one of competing events. More precisely, it denotes the instantaneous rate of occurrence of the c event in subjects who are currently event free (*i.e.*, in subjects who have not yet experienced any of the different types of events). The Cause-specific hazard function for event type c can be expressed as:

$$h_c(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T_c < t + \Delta t | T_c \geq t)}{\Delta t}$$

T_c indicates the time to failure from event type c . Each hazard ratio represents the instantaneous relative risk of an event of interest in the presence of the covariate (e.g., the ratio of the hazard rates corresponding to the conditions described by two different levels of an explanatory variable, all other covariates being equal). However, this hazard ratio cannot be directly translated to the cumulative incidence function which is clinically relevant and may provide useful information to researchers. The Fine-Gray model addresses this issue and has the advantage that the cumulative incidence of the event of interest has a direct link with the estimated sub-distribution hazard, and thus regression coefficients quantify the direct effects of covariates on the cumulative incidence [41]. Instead, cause-specific hazard models may be more appropriate for addressing questions about the causes or origins of a disease [55].

Fine-Gray Subdistribution hazard model Fine and Gray (1999) proposed a proportional hazards model to estimate the effect of covariates on the cumulative incidence function for the event of interest treating the CIF curve as a subdistribution function. The Fine and Gray subdistribution hazard function for event type c can be expressed as:

$$h_{c,CIF}(t) = \lim_{\Delta t \rightarrow 0} \frac{Pr(t < T_c < t + \Delta t | T_c > t \cup T_{c'} \leq t, c' \neq c)}{\Delta t}$$

It denotes the instantaneous risk of failure from the c event in subjects who have not yet experienced an event of type c . Note that this risk set includes those who are currently event free as well as those who have previously experienced a competing event. This differs from the risk set for the cause-specific hazard function, which only includes those who are currently event free.

Using the same example as above, the subdistribution hazard of cardiovascular death denotes the instantaneous rate of cardiovascular death in subjects who are still alive (*i.e.*, who have not yet experienced either event) or who have previously died of noncardiovascular causes.

6.0.2 Performance metrics

6.0.2.1 C-index

The concordance index or C-index is a goodness of fit measure to evaluate the global discrimination power of a survival model. In other words, it measures the model's ability to correctly provide a reliable ranking of the survival times based on the individual risk scores where data may be censored. It is computed from the risk scores and times-to-event data of pairs of subjects. The intuition behind the formula is the following:

$$\text{C-index} = \frac{\# \text{ concordant pairs}}{\# \text{ concordant pairs} + \# \text{ discordant pairs}}$$

Given a pair of non-censored subjects (i, j) , it is a concordant pair if: $\eta_i > \eta_j$ and $T_i < T_j$. It is consider a discordant pair if $\eta_i > \eta_j$ and $T_i > T_j$. If one of T_i and T_j is censored, the subject that is not censored is observed. For instance, if T_j is censored, if $T_j < T_i$ the pair is discarded since it can not be established which subject got the disease first. Otherwise, if $T_j > T_i$, then we can inspect risk score condition as before. In the case that both subjects are censored, the pair is discarded. This logic is represented in the following formula:

$$\text{C-index} = \frac{\sum_{i,j} 1_{T_j < T_i} \cdot 1_{\eta_j > \eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j < T_i} \cdot \delta_j}$$

where η_i is the risk score of the subject i .

- $1_{T_j < T_i} = 1$ if $T_j < T_i$ else 0

- $1_{\eta_j > \eta_i} = 1$ if $\eta_j > \eta_i$ else 0

A C-index=1 corresponds to the best model prediction, C-index=0.5 represents a random prediction and C-index=0 means that the model does not have discrimination power at all.

6.0.2.2 Brier score

The Brier score is a measure used to evaluate the accuracy of a predicted survival function at a given time t for binary outcome events. It is represented by the average squared distances between the observed survival status and the predicted survival probability.

$$BS = \frac{1}{N} \sum_{t=1}^N (f_t - o_t)^2$$

where N is the total number of subjects, f_t is the predicted probability value for time t and o_t is the value of the outcome event.

The output score is always a number between 0 and 1, with 0 being the best possible value. In terms of benchmarks, a useful model will have a Brier score below 0.25.

6.0.3 Approaches

Survival analysis has the purpose of analyzing in-hospital mortality in order to estimate and interpret survival and hazard functions from the survival data and to assess the relationship of explanatory variables to survival time. In-hospital mortality is derived from general hospital mortality. Registered mortality can have five values: hospital discharge alive, *exitus letalis*, transferred to another hospital, transferred to a health center, or unknown (this patients are discarded in the current study). For those patients that have been transferred, they are consider alive if and only if the center to which they have been transferred is less complex. In other words, the patient has improved and does not need to stay in the hospital but should continue to be under regular medical supervision. On the contrary, those patients for whom this cannot be assured are considered non-analyzable and therefore excluded from the study. Fortunately, this phenomenon is observed in less than 8% of total patients.

Prior to the analysis, a feature analysis has been carried out using survival forests. The results of the feature analysis can be seen in 7. Two ways to proceed have been defined, they are named approach A and approach B and are defined below:

Approach A First, this approach classifies patients between survivors and non-survivors through the use of linear classifiers. Then, survival models are developed only with non-surviving patients (443) modeling the hospital death event and using time in hospital as a time to event column. In this case, the survival function is only examined for non-surviving patients. The whole process is depicted in Figure 6.2.

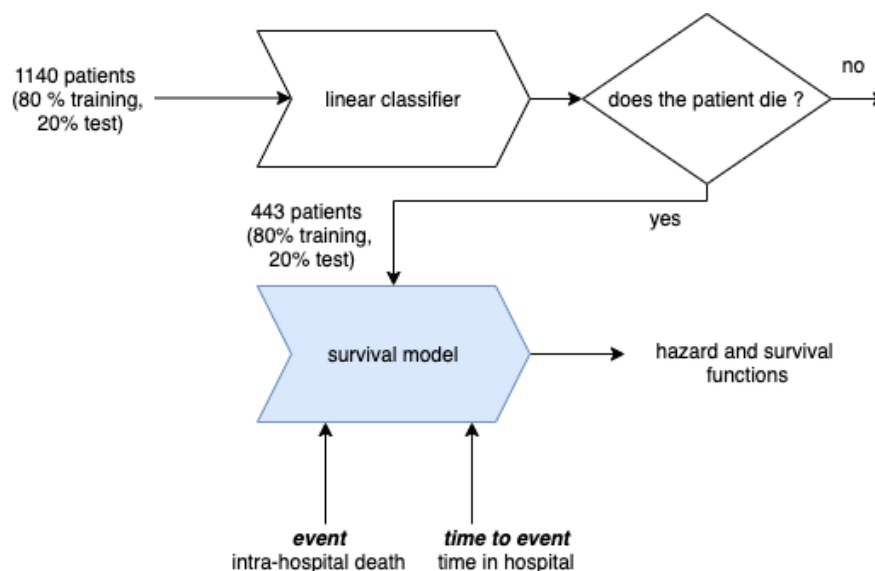


Figure 6.2: Pipeline of approach A

Approach B This procedure works in reverse of the previous one. Survival models are first defined with all patients using ICU discharge as event and time in ICU as time to event column. Afterwards, it is determined whether the ICU discharge was motivated by an improvement or by death using linear classifiers. Those classifiers receive as input the survival probabilities computed by survival models apart from receiving patient history. The reason why it was decided to model the ICU discharge event is to be able to use the information of all the patients instead of using only non-survivors. Since all patients are discharged from ICU for one reason or another, we can use all of them as input for the survival model. The pipeline of this approach can be seen in Figure 6.3.

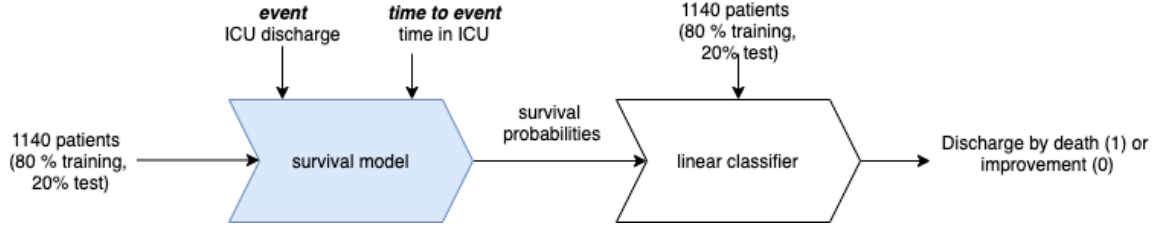


Figure 6.3: Pipeline of approach B

In addition to traditional survival models, competitive risk models have also been used. The fact that leaving ICU due to improvement constitutes a competitive event since it makes it impossible to leave the ICU due to death, makes the use of models for competitive risks convenient.

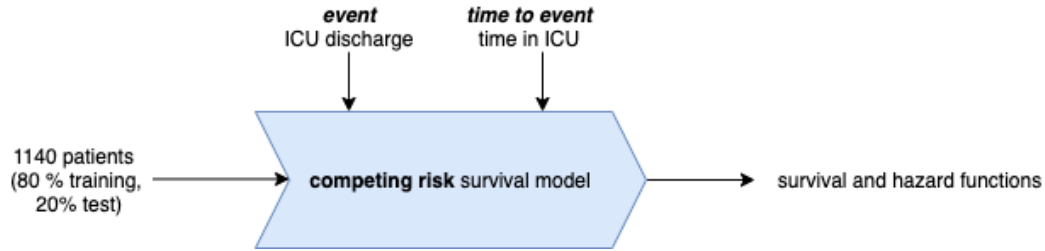


Figure 6.4: Pipeline of approach B using competing risk survival models

Other analyses and methods In an exploratory fashion, this research also includes variants of approaches A and B (excluding competitive risk models). We ignore date and time variables in order to analyze whether it is possible to obtain good performance using only variables available at the beginning of ICU admission. Thus, variables such as *alive28Days*, time in ICU and time in hospital are excluded when possible (they cannot be excluded if they act as a time to event column). In contrast, variables that had not been used in previous survival models but were used in the imputation of missing values have been included. In addition, approaches A and B were replicated removing correlated variables. For more information visit 7.

Additionally, previous work is also included using Cox's proportional hazards model, Cox's time varying proportional hazards and non-linear Cox models. Nevertheless, they did not achieve clinically acceptable results. This led us to consider the problem from another perspective.

The results of feature selection analysis along with those of survival analyses for each procedure are evaluated in depth in Chapter 7.

Chapter 7

Results

In this section we present the results of the survival analysis. First we show the results of the A (7.2) and B (7.3, 7.4) approaches and then previous and complementary analyses. As mentioned in previous sections, before carrying out any of these approaches, we have performed feature analysis using survival forests. The importance of each feature is evaluated using Conditional Survival Forest (CSF), Random Survival Forest (RSF) and Extra Survival Trees (EST). The idea is to discover which set of features is better when describing the event that we want to model as if it were dealing with a simple classification. The importance of each variable is stored after training in a dictionary, that is a default attribute of all survival forest models. The quality of the analysis is then tested in the prediction task.

We decided to use survival forests in contrast to traditional machine learning techniques to be able to analyze survival with sets of features different from those obtained in 4.1.8. In addition these methods designed especially for survival data, have advantages over common problems such as overfitting, inflated standard errors and convergence problems.

Although they have several attributes to customize, we have only experimented with the number of features randomly chosen at each split (*max_features*). The reason for this is that inspecting the effect that each of the attributes have on the quality of the feature selection is not part of the objectives of this project.

To evaluate the results of feature and survival analysis, the concordance index and brier score metrics previously presented in 6.0.2 are used on the testing dataset. Linear classifiers have been assessed using 10-fold cross validation and evaluated using the accuracy, precision, recall and f1-score metrics. As linear classifiers we have used SVM (linear and kernel based when convenient), logistic regression and perceptron.

The implementations of these ensemble models and those used in survival analysis have been extracted from the *PySurvival* [30] and *lifelines* [57] library while *Sklearn*[29] library has been used to test the linear classifications.

7.1 Variable selection

Before the implementation of feature selection and approaches A and B, a specific subset of variables was first defined from which later the final selection is then carried out using survival forests. In the tables 7.1, 7.2 the list of variables from which the survival analysis starts is shown. This selection has been chosen by medical professionals based on the clinical weight of the variables and risk indicators demonstrated in other studies. Some variables were eliminated because they were not clinically relevant or because they were variables that were collected only in few hospitals.

Topic	Metrics
Complications	pneumotorax, pleural effusion, organizing pneumonia, pulmonary embolism, cardiac arrest, bacteremia, ictus, dic, chronic kidney disease, hepatic dysfunction, hemorrhage, hyperglycemia, hypoglycemia, lung coinfection
Outcomes	outcome simplified (alive or death) exitus date, outcome (hospital discharge alive, exitus lethalis, transferred to another hospital, transferred to a social health center, unknown), alive28Days

Table 7.2: Initial features selected for survival analysis of adults admitted with COVID-19 [1/2]

Topic	Metrics
Patient characteristics	age, gender
Previous Medication	aceInhibitors, arb, beta blockers, calcium channel blockers, diuretics, statins, streptococcus vaccine, flu vaccine
Comorbidities	hypertension, obesity, diabetes, hematological comorbidities, renal chronic disease, asthma, hiv, heart chronic disease, pulmonar chronic disease, dementia, severe hepatic disease
Laboratories (only ICU and 3rd day ICU)	il6, requirement of vasopressors, requirement of ECMO, breathing rate, heart rate, temperature, HCO ₃ , prothrombin time, total bilirubin, ALT, AST, troponin-I, troponin-T, NT-proBNP, paCO ₂ , LDH, leucocytes, creatinine, lymphocytes, CRP, procalcitonin, lactate, D-dimer, ferritin, urea, platelets, glucose, immunodeficiency
Mechanical Ventilation (only ICU and 3rd day ICU)	ventilatory ratio, ventilatory ratio <i>modified</i> , IMV requirement, NIMV requirement, compliance, paFi, oxygen saturation, previous respiratory support, RASS, SAS, driving pressure, plateau pressure PEEP pressure
Hospital Course	hospital, symptoms start date, hospital admission date, discharge date, time in hospital, ICU admission date, ICU discharge date, time in ICU, start date of IMV end date of IMV, time receiving IMV, time between hospital admission, symptoms start date
Complementary Therapies	prono position needed, recruitment manouvers, neuromuscular blockers
Treatments	corticosteroids, antibiotics, lopinavir/ritonavir, hidroxicloroquina, tocilizumab, interferon beta

Table 7.1: Initial features selected for survival analysis of adults admitted with COVID-19 [2/2]

7.2 Approach A

In this approach we use the time in hospital and the rest of the variables to model in-hospital death. We have 433 non-surviving patients, 80% (348) of which are used for training and the remaining 20% (88) for testing survival forest models. In feature selection, for each survival forest we try three different configurations for the variable *max_features* where N is the total number of variables, in each split we can consider taking into account: \sqrt{N} , $\log_2(N)$ or N features. Other parameters such as the number of trees (*num_trees* = 100) used or the split criterion (*importance_mode* = *impurity_corrected*) are kept constant for all trees.

In Table 7.3, feature selection results are shown. For each survival forest and classification performance, results are shown according to concordance index and brier score. The methods are listed in ascending order according to the concordance index. It is observed that C-index is higher than 0.5 for all the trees. Recall that the higher the value of this index, the greater the discriminatory power of the model, with 1 being the perfect discriminatory power. Based on this, we can affirm that all models are good at differentiating between high and low risk patients.

Brier score values are also acceptable: they are less than 0.25. A Brier score of 0 reflects perfect accuracy (*i.e.*, there is no difference between event scores (in this case in-hospital death) and someone's probabilistic predictions for those events), and a Brier score of 1 reflects perfect inaccuracy (*i.e.*, events that not occur receives probabilities of 1 while events that do occur receives probabilities of 0).

Generally, methods with *max_features* = *all* perform better. This may be due to the fact that all features are taken into account at the time of split while using other settings, important features can be omitted. Regarding the selection of features for *log2* or *sqrt*, no performance differences are observed. The best results are achieved by CSF and RSF (C-index=0.90, Brier score=0.03).

The variables that appear in the top 5 of most methods are: time in IMV (9/9) time in ICU (7/9), alive28days (7/9) infection complications lungs (6/9) and bacteremia (5/9). The feature set provided by CSF (*max_features* = *all*) has been chosen for survival analysis, shown in Table 7.4. The variables in the table do not follow any particular order. Among the variables with importance, none belonging to patient characteristics, treatments or complementary therapies have been selected. The hypothesis that is considered that could be the cause of this phenomena is that the variables at the top are sufficiently descriptive. This is the case of alive28Days, which indicates whether a patient has remained 28 days alive. This variable has been removed from the selection

method	C-index	brier score
CSF (<i>max_features = all</i>)	0.90	0.03
RSF (<i>max_features = all</i>)	0.90	0.03
EST (<i>max_features = all</i>)	0.89	0.04
RSF (<i>max_features = sqrt</i>)	0.86	0.06
RSF (<i>max_features = log_2</i>)	0.85	0.06
CSF (<i>max_features = sqrt</i>)	0.84	0.06
EST (<i>max_features = log_2</i>)	0.80	0.07
EST (<i>max_features = sqrt</i>)	0.79	0.06
CSF (<i>max_features = log_2</i>)	0.79	0.07

Table 7.3: Feature selection results for approach A

because it has a high correlation with the in-hospital death event variable, since excepting 17.46% of patients, the rest have not survived the 28 days.

7.2.0.1 Classification according to survival

Once the features have been selected, the next step is to train several linear classifiers to predict whether a patient will survive or not based on the set of features. From the 1,140 patients, four different classifiers have been trained and subsequently validated using 10-fold cross validation. The best model has been selected based on the f1-score. For each algorithm, in Table 7.5 we show the precision, recall, f1-score and accuracy metrics. Although the four algorithms obtain similar accuracy-based performances, the sensitivity is slightly worse for logistic regression and perceptron.

In order to fully understand the particularities of each classifier and to know which features have been most relevant when classifying patients between survivors and non-survivors, methods from the ELI5 library [58] have been used for explainability purposes.

The methods available for lineal classifiers consist of explaining the weights of the variables in the algorithm and in specific predictions. Figures 10.123, 10.125 and 10.124 show the resulting weights for lineal SVM, logistic regression and perceptron respectively. Weights can be either positive or negative according to it's sign. Positive values correlate with growing chance of classifying as positive class (non-survivors), and negative ones with chance becoming negative samples (survivors). In other words, the sign of the feature shows towards which class a feature is more correlated to. The feature that appears as ̢BIAS_i refers to the intercept term. It determines where the separation line

Topic	Metrics
Patient characteristics	-
Previous Medication	statins, flu vaccine, beta blockers, interferon beta, diuretics, ACE inhibitors
Comorbidities	hematological disease, obesity, renal chronic disease
Laboratories (only ICU and 3rd day ICU)	CRP (ICU, 3r day), septic shock (ICU), platelets (ICU), SOFA (3r day), glucose (ICU), leucocytes (3r day), lymphocytes (ICU), lactate (ICU), creatinine (ICU, 3r day), procalcitonin (ICU)
Mechanical Ventilation (only ICU and 3rd day ICU)	ventilatory ratio (ICU, 3rd day), previous respiratory support (non-invasive ventilation helmet), paFi (3r day), oxygen saturation (ICU)
Hospital Course	time in ICU, time in IVM, days between hospital admission and symptoms started, ICU and IMV times equal
Complementary Therapies	-
Treatments	-
Complications	bacteremia, pneumotorax, infectious complications: lungs, hemorrhage
Outcomes	alive28Days

Table 7.4: Feature set provided by CSF (*max_features = all*)

method	class	precision	recall	f1-score	accuracy
SVM linear kernel	no-survivor	0.97	0.94	0.95	0.96
	survivor	0.95	0.98	0.96	
SVM RBF kernel	no-survivor	0.95	0.93	0.94	0.95
	survivor	0.95	0.96	0.96	
logistic regression	no-survivor	0.97	0.89	0.93	0.95
	survivor	0.94	0.98	0.96	
perceptron	no-survivor	0.93	0.88	0.91	0.94
	survivor	0.95	0.97	0.96	

Table 7.5: Performance of lineal classifiers in approach A

intercepts the y-axis, although it can also be seen as the offset that is added to all predictions (reason for which is called bias in the machine learning field).

In general, we observe that the features showing higher correlations for non-survivors are *equal_ICU_IMV_times*, complications (renal chronic disease, hemorrhage, pneumothorax, bacteremia), patient prognostic scores (SOFA score on the 3rd day) and previous medication (statins). While for survivors we find laboratory variables (platelets, oxygen saturation and creatinine for ICU event), obesity and BIAS.

These results show some behaviors already detected in the statistical analysis. The times in ICU and IMV are much longer for the survivors, the complications reported by the classifiers present differences in important percentages and the sofa on the third day is higher for non-survivors. However, it is not clear why statins appear at the top and another medication such as the flu vaccine does not (also with a large percentage difference). Regarding the survivors, obesity appears with a high correlation just after the BIAS. This may be caused by the nature of the samples since it was observed that 36.49% of the survivors were obese compared to 20.19% of non-survivors.

7.2.0.2 Survival analysis

Once we know if the patient in question is going to be part of the group of non-survivors, the next thing we want to know is when that patient is going to die. That is, the time that will elapse until the hospital death event takes place.

Kaplan-Meier is one of the simplest methods for describing survival of a study population and to compare two study populations. Together with the Cox model, they constitute the conventional methods of survival analysis. Fig-

Figure 7.1 shows the percentage of survivor patients in each time step. This curve indicates that after 16-18 days, the 50% of population has already experienced the outcome event (in-hospital death) and that after 60 days the number of survivors is close to 0. Time t_0 represents the chance in survival right on the first day of hospital admission.

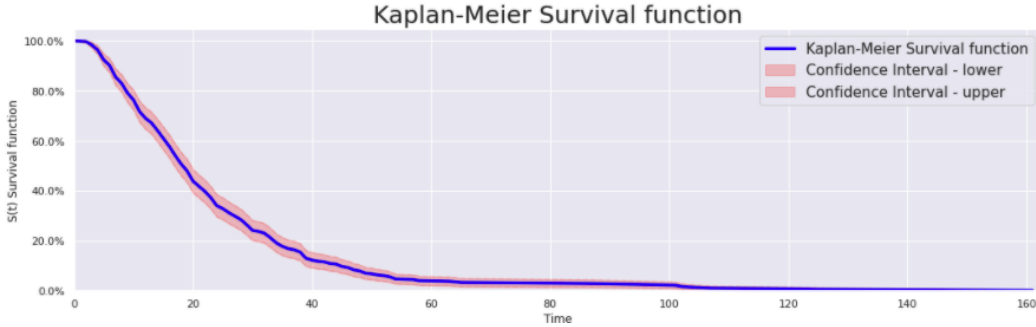


Figure 7.1: Kaplan Meier survival curve for approach A

Apart from using Kaplan-Meier, we have used other survival models such as Conditional Survival Forest (CSF), neural SVM, Cox Proportional Hazards model (Standard CoxPH) and non-linear Cox. For non-linear Cox, we performed several experiments with different hyperparameters (mainly activation function, number of units and dropouts). However, the best performance was achieved by a two-layer architecture with ReLu, 128 units per layer.

They were trained only with data from non-survivor patients, *i.e.*, 348 patients for training and 88 for testing. In Table 7.6 we present the performance results of these models based on Concordance index and Brier score. All the models present a fairly similar performance although the CSF once again stands out with the best as we saw in the feature selection process.

The following 7.2, 7.3 and 7.4 Figures show the inferred survival curves of the CSF, non-linear cox and Standard CoxPH models respectively for a specific sample. The sample corresponds to a patient who died 20 days after being admitted to the ICU. This is indicated by a vertical line at 20 days on all charts.

The survival curves show for each time interval, in our case the number of days hospitalized, the probability of survival. Although all curves may appear to be the same at first glance, there are notable differences between them. The CSF and non-linear Cox curves begin to decrease from approximately the tenth day, while the Cox curve decreases already on the first day. This places the patient at risk in advance. On the other hand, the slope of the curve is much steeper for non-linear Cox, which leads to a much shorter life

method	C-index	brier score
CSF (max_features='all', num_trees=200)	0.89	0.05
neural SVM	0.89	-
Standard CoxPH	0.87	0.04
non-linear Cox	0.86	0.04

Table 7.6: Performance of survival models in approach A

expectancy than in the other models. For CSF and Cox we observe that the models stabilize from approximately 30-35 days while for non-linear this occurs after 20 days. Taking this into account, the probabilities of survival for this patient on day 20 are approximately 0.5 (CSF), 0.0 (non-linear Cox) and 0.25 (Standard CoxPH).

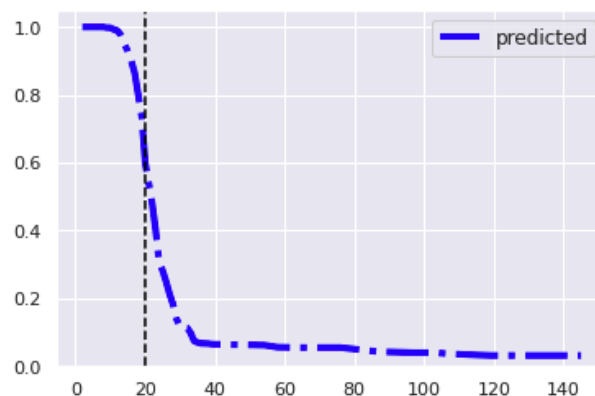


Figure 7.2: Sample survival curve produced by CSF

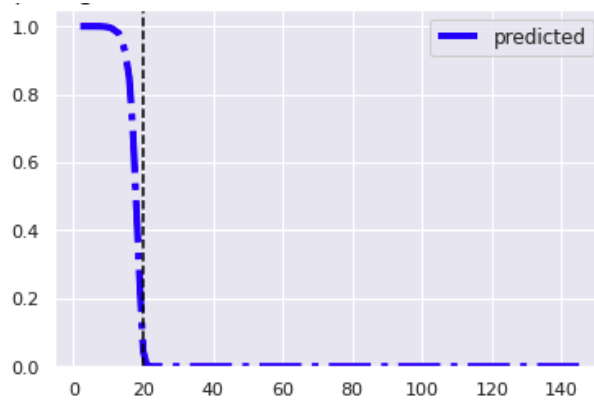


Figure 7.3: Sample survival curve produced by non-linear Cox

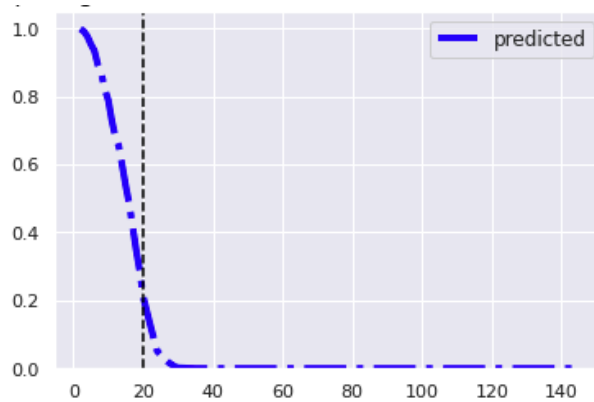


Figure 7.4: Sample survival curve produced by CoxPH

In addition, the Cox model performs a multivariate regression shown in Figures 7.5 and 7.6. For each feature, the Cox model computes the statistical significance (z), the regression coefficients (*coeff*), the standard error (*std.err*), upper and lower confidence intervals (*lower.ci*, *upper.ci*) and p-values (chi squared test). The hazard ratios are shown in Table 7.7, significant variables are in bold.

variables	coef	std. err	lower_ci	upper_ci	z	p_values
timeInICU	-1.237	0.014	-1.265	-1.209	-87.387	0.0
timeInIMV	-0.666	0.013	-0.692	-0.641	-50.443	0.0
crp_ICU	0.084	0.005	0.074	0.094	16.739	0.0
septic_shock_ICU	-0.187	0.182	-0.543	0.17	-1.025	0.305
ventilatoryRatio_ICU	-0.121	0.096	-0.309	0.068	-1.254	0.21
equal_ICU_IMV_times	0.265	0.232	-0.189	0.719	1.143	0.253
bacteremia	-0.063	0.136	-0.33	0.203	-0.466	0.641
ventilatoryRatio_day3	-0.038	0.082	-0.199	0.122	-0.467	0.64
platelets_ICU	0.006	0.001	0.005	0.007	10.753	0.0
Hospital Admission-Start Symptoms	0.233	0.01	0.214	0.252	24.218	0.0
sofa_day3	0.313	0.023	0.269	0.358	13.841	0.0
pneumotorax	0.082	0.177	-0.264	0.428	0.464	0.643
glucose_ICU	0.046	0.001	0.044	0.047	61.2	0.0
statins	0.019	0.137	-0.25	0.287	0.135	0.893
fluVaccine	0.009	0.162	-0.309	0.327	0.057	0.955
betaBlockers	0.014	0.196	-0.37	0.397	0.069	0.945
PaFi_day3	-0.122	0.001	-0.124	-0.121	-137.415	0.0

Figure 7.5: Multivariable Cox [1/2]

PaFi_day3	-0.122	0.001	-0.124	-0.121	-137.415	0.0
Infectious complications: Lungs	-0.031	0.136	-0.297	0.236	-0.227	0.82
Interferon beta	0.105	0.13	-0.15	0.361	0.808	0.419
hematological	-0.231	0.251	-0.723	0.26	-0.922	0.357
obesity	0.134	0.132	-0.125	0.393	1.015	0.31
renalChronic	0.045	0.256	-0.457	0.547	0.174	0.862
diuretics	0.105	0.175	-0.239	0.449	0.596	0.551
leucocytes_day3	-0.136	0.012	-0.158	-0.113	-11.678	0.0
hemorrhage	-0.015	0.177	-0.361	0.331	-0.082	0.935
lymphocytes_ICU	0.206	0.065	0.079	0.333	3.185	0.001
crp_day3	0.208	0.005	0.198	0.217	41.106	0.0
acelnhibitors	-0.028	0.162	-0.345	0.289	-0.174	0.862
lactate_ICU	0.078	0.005	0.069	0.088	16.105	0.0
creatinine_day3	0.022	0.062	-0.1	0.144	0.356	0.722
creatinine_ICU	-0.03	0.115	-0.255	0.194	-0.265	0.791
procalcitonin_ICU	-0.007	0.009	-0.025	0.011	-0.774	0.439
variablesOxygenSaturation_ICU	-0.033	0.008	-0.048	-0.018	-4.294	0.0
previousRespiratorySupport, Casco de ventilaci...	-0.03	1.033	-2.054	1.994	-0.029	0.977

Figure 7.6: Multivariable Cox [2/2]

Among the statistically significant variables, we find CRP (ICU, 3rd ICU day), SOFA score (3rd ICU day), presence of symptoms and lymphocytes (ICU) associated with a higher risk of death, and platelets (ICU), glucose (ICU), lactate (ICU) to a lesser extent. As variables associated with a good prognosis we find *timeInICU*, *timeInIMV*, paFi (day3) and leukocytes (day3). Longer times of hospitalization or IMV does not have a negative impact on life expectancy and as paFi increases, it can be considered that the quality of the patient's breathing improves.

Recall that with a continuous variable, the hazard ratio indicates the change in the risk of death if the parameter in question rises by one unit. For example, for our model lymphocytes (ICU) has associated a HR of 1.228 and leukocytes (3rd ICU day) a HR of 0.872. For lymphocytes (ICU), given one patient in the ICU for each unit that the lymphocytes increase, the higher the risk of death is considered ($HR > 1$), specifically the risk is 22.8 %. While for leukocytes (3rd ICU day) for each unit the risk of death decreases 12.8 % ($HR < 1$). However, this phenomena is not actually observed in real situation,

and statistical analysis showed that the number of leukocytes for non-survivors is higher 5. Although leukocytes always show a non-linear behavior, u-shaped to be exact, these values are not commonly seen. For this reason, in the future we consider to build a restricted cubic spline to model the relationship between leukocytes and the outcome variable and discover what is currently happening with leukocytes at 3rd day of ICU.

Variable	coeff	exp(coeff)
timeInICU	-1.237	0.290
timeInIMV	-0.666	0.513
crp_ICU	0.084	1.087
septic_shock_ICU	-0.187	0.829
ventilatoryRatio_ICU	-0.121	0.886
equal_ICU_IMV_times	0.265	1.30
bacteremia	-0.063	0.938
ventilatoryRatio_day3	-0.038	0.962
platelets_ICU	0.006	1.006
Hospital Admission-Start Symptoms	0.233	1.262
sofa_day3	0.313	1.367
pneumotorax	0.082	1.085
glucose_ICU	0.046	1.047
statins	0.019	1.019
fluVaccine	0.009	1.009
betaBlockers	0.014	1.014
paFi_day3	-0.122	0.885
Infectious complications: lungs	-0.031	0.969
Interferon beta	0.105	1.11
hematological	-2.231	0.107
obesity	0.134	1.143
renalChronic	0.045	1.046
diuretics	0.105	1.110
leucocytes_day3	-0.136	0.872
hemorrhage	-0.015	0.985
lymphocytes_ICU	0.206	1.228
crp_day3	0.208	1.231
aceinhibitors	-0.028	0.972
lactate_ICU	0.078	1.081
creatinine_day3	0.022	1.022
creatinine_ICU	-0.03	0.970
procalcitonin_ICU	-0.007	0.993
variablesOxygenSaturation_ICU	-0.033	0.967
previousRespiratorySupport, casco de ventilación no invasiva	-0.03	0.970

Table 7.7: Coefficient and hazard ratios for Cox model in approach A

7.3 Approach B

In this approach we use as column time the time spent in ICU and the event ICU discharge to model death in ICU. Since all the patients were admitted in ICU, we can feed the models using the whole dataset. However, patients that did not have specified time in ICU or admission and discharge dates were ignored in survival analysis. In total, we obtained 1129 patients.

The analysis of features is analogous to the analysis of approach A. The same models and configurations have been used. Results are shown in 7.8. The results are sorted according to the C-index. Again we observe that the methods with *max_features = all* have better performances. Although both C-index and Brier score present acceptable values at the clinical level, the C-index is slightly lower compared to the previous approach.

The variables that appear in the top 5 of most methods are the ones seen in previous approach. These variables influence both ICU discharge and hospital discharge, either due to death or improvement. The feature set provided by RSF (*max_features = all*) has been chosen for survival analysis, shown in Table 7.9.

In this case, we see that new variables appear. In patient characteristics we have now the age. The number of prior medication variables is reduced and tocilizumab is included. The set of comorbidities is completely different. The laboratory variables are quite similar but more importance is given to the 3rd day at ICU event and the set of ventilation variables is reduced. We also found treatment and complementary therapies variables. The hospital course variables remain the same. For outcomes, alive28Days reappears but is not taken into account in the survival analysis.

7.3.0.1 Survival analysis

Firt of all, we compute the Kaplan-Meier estimator for the population of 1,129 patients. Figure 7.7 shows the survival function for the population of which the ICU discharge event was recorded. The curve is very similar to previous one obtained in approach A. After 16-18 days, half of the patients have already left the ICU and after 80 days all have been discharged.

method	C-index	brier score
RSF(max_features="all")	0.85	0.04
CSF (max_features="all")	0.83	0.04
EST (max_features="all")	0.83	0.04
RSF (max_features="sqrt")	0.82	0.05
CSF (max_features="sqrt")	0.80	0.06
EST (max_features="sqrt")	0.79	0.06
CSF (max_features="log_2")	0.79	0.07
RSF (max_features="log_2")	0.78	0.06
EST (max_features="log_2")	0.76	0.06

Table 7.8: Feature selection results for approach B

Topic	Metrics
Patient characteristics	age
Previous Medication	statins, interferon beta, tocilizumab
Comorbidities	hepatic dysfunction, acute kidney failure, diabetes, flu vaccine
Laboratories (only ICU and 3rd day ICU)	urea (3r day), glucose (ICU, 3r day), CRP (ICU, 3r day), lymphocytes (ICU), creatinine (ICU, 3r day), septic shock (3r day), platelets (ICU, 3r day), hemodynamic SOFA(ICU, 3r day), leucocytes (3r day), LDH (ICU), procalcitonin (ICU)
Mechanical Ventilation (only ICU and 3rd day of ICU)	paFi (ICU), ventilatoryRatio (3r day), previous respiratory support oxygen mask with reservoir bag
Hospital Course	time in ICU, time in IVM, days between hospital admission and symptoms started, ICU and IMV times equal
Complementary Therapies	neuromuscular blockers, recruitment manouvers
Treatments	corticosteroids, lopinavir/ritonavir
Complications	infectious complications: lungs, bacteremia
Outcomes	alive28Days

Table 7.9: Feature set provided by RSF (*max_features = all*)

approach	C-index	brier score
non-linear Cox	0.86	0.04
neural SVM	0.84	-
standard CoxPH	0.80	0.04
CSF (max_features='all', num_trees=200)	0.69	0.06

Table 7.10: Performance of survival models in approach B

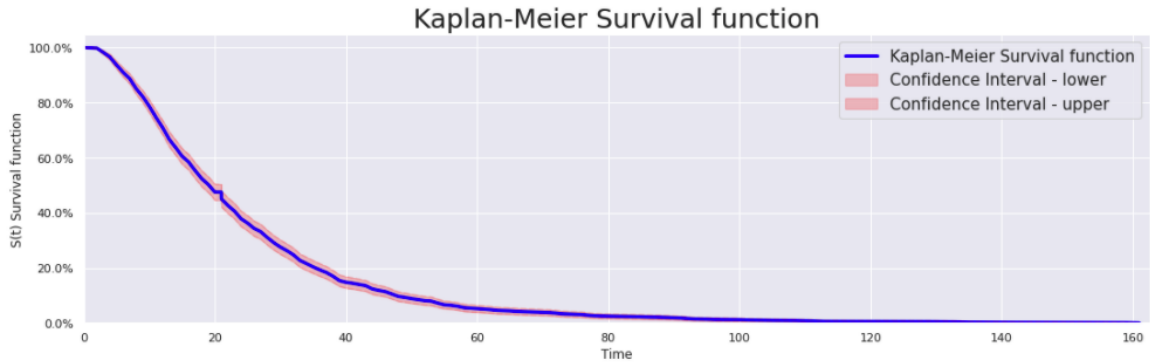


Figure 7.7: Kaplan Meier survival curve for approach B

Afterwards, we have used other survival methods to model survival probabilities. The same models as in approach A have been used, the results can be seen in Table 7.10.

This time we see that the performances are reversed. The method that obtains a higher C-index is the non-linear Cox while the CSF obtains an index lower than 0.70. The corresponding survival curves for a specific patient are shown in Figures 7.8, 7.9 and 7.10. The sample corresponds to a patient who was discharged after 34 days in ICU. Although CSF has quite low yields, the curve is more representative than the one shown by non-linear Cox. This last curve is advanced and confirms that the patient is discharged earlier, approximately at the 25th day. The Cox curve is similar to CSF's curve but with a smoother slope, at 34th day the patient has a probability of approximately 25% of remaining in the ICU.

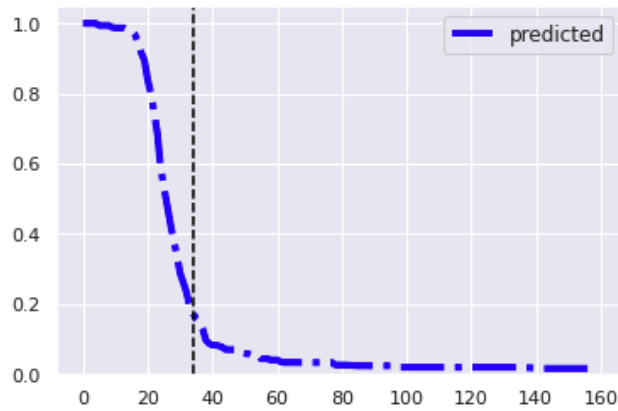


Figure 7.8: Survival curve produced by CSF

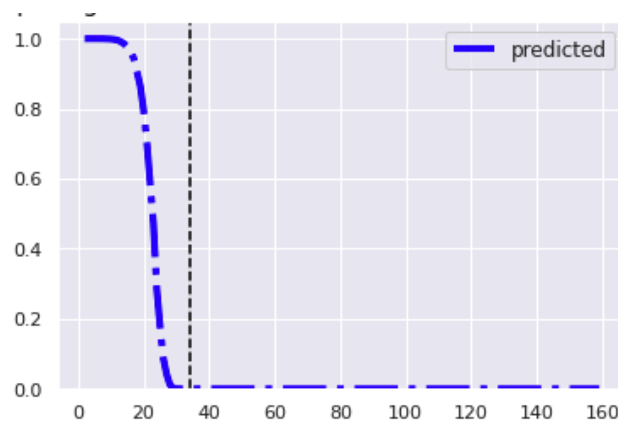


Figure 7.9: Survival curve produced by non-linear Cox

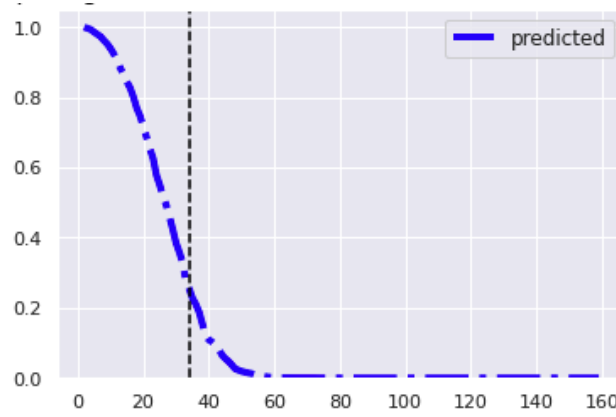


Figure 7.10: Survival curve produced by CoxPH

The resulting survival probabilities of each patient are included in the dataset of the 1,129 patients to train several classifiers and determine if the ICU discharged is caused by an improvement or death. The performance results are shown in Figure 7.11. The models used are again SVM, logistic regression and perceptron. For exploratory purposes, this time we have tested SVM with polynomial, sigmoid and RBF kernel.

Sorted according to accuracy, the models achieving higher C-index are SVM with linear and polynomial kernel. In third place we find the logistic regression, the level of accuracy is penalized by slightly low precision when classifying non-survivors. The following models stand out for having a fairly poor precision classifying survivors and quite low sensitivity levels for classifying non-survivors.

Next, we analyze the weights of the models in Figures 10.128, 10.129 and 10.130. SVM with polynomial, sigmoid and RBF are ignore since ELI5 only has methods for linear SVM. At a general level, it can be seen that *equal_ICU_IMV_times*, *timeInIMV* are highly correlated with ICU discharge in all the models. We also find comorbidities (acute kidney failure), treatments (corticosteroids), complications (bacteremia, lung infectious complications), laboratory variables for 3rd day (CRP and leucocytes) and age. Age appears as the feature with the greatest weight in the perceptron.

method	class	precision	recall	f1-score	accuracy
SVM linear kernel	no-survivor	0.96	0.93	0.95	0.97
	survivor	0.97	0.98	0.98	
SVM polynomic kernel	no-survivor	0.97	0.89	0.93	0.95
	survivor	0.94	0.98	0.96	
logistic regression	no-survivor	0.87	0.91	0.89	0.91
	survivor	0.93	0.90	0.92	
perceptron	no-survivor	0.84	0.36	0.50	0.72
	survivor	0.69	0.95	0.80	
SVM RBF kernel	no-survivor	0.84	0.26	0.40	0.71
	survivor	0.70	0.97	0.81	
SVM sigmoid kernel	no-survivor	0.67	0.44	0.53	0.68
	survivor	0.68	0.84	0.75	

Table 7.11: Performance of lineal classifiers in approach B

7.4 Competing risk survival models

In this section we model the ICU discharge event with competitive risk models. As previously stated, the ICU discharge event due to improvement is a competitive event since it prevents us from observing the ICU death event for that patient. The methods that have been used are: CIF for estimating the percentage of survivors in each time unit, cause-specific hazard model and subdistribution hazard model for estimating influence of covariates on the event according to each type of risk. However, recall that for this purpose is better to use subdistribution hazard model. Cause-specific hazard models are more appropriate for addressing questions about the causes or origins of a disease.

All these methods are implemented in *R* [59]. CIFs can be estimated using the *cuminc* function in the *cmprsk* package. Cause-specific hazard models can be fit using the *coxph* function in the *survival* package treating those subjects who experience a competing risk as being censored at the time of the occurrence of the competing event. Subdistribution hazard models can be fit in *R* by using the *crr* function in the *cmprsk* package as well.

The same features have been used as in approach B (1,129 patients). In addition, a new variable called *ICU_discharge_by_death* has been added to distinguish if ICU discharge has been due to improvement or death. Cumulative incidences of ICU discharge by improvement and by death in the overall

sample are described in Figure 7.11. The proportion of subjects who experience the event of death is lower than those who leave the ICU due to improvement. In the statistical analysis we saw that 38.86% of the patients died, in this case our sample is 1,129 so this percentage is slightly lower. After approximately 50 days, almost all patients in the non-survivors cohort have died. While for the survivors, they tend to spend more time in the ICU. After 100 days almost all the surviving patients have been discharged.

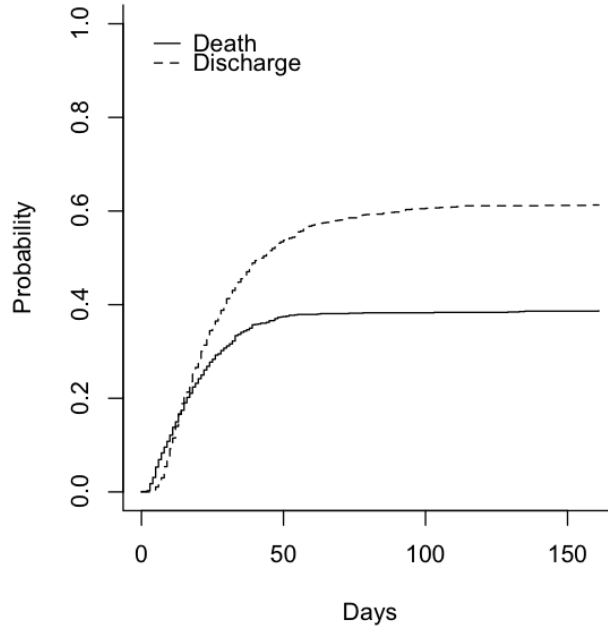


Figure 7.11: Crude incidence of ICU discharge by improvement and by death

We fit cause-specific and subdistribution hazard models for both ICU discharge improvement and death. The estimated hazard ratios, along with their statistical significance are reported in Tables 7.12, 7.13 for Subdistribution hazard model and 7.14, 7.15 for Cause-specific hazard model when patients are discharged by death.

Hazard ratios above 1 indicates that the covariate is positively associated with the event probability, that is ICU discharge by death. This is why covariates with hazard ratios > 1 are called bad prognostic factors. Among the statistically significant variables, we observed with positive hazard ratios in 7.12, 7.13: *equal_ICU_IMV_times*, age, glucose (3rd day), lymphocytes (ICU), acute kidney failure, neuromuscular blockers requirement, hemodynamic SOFA (3rd day). These variables act as risk factors that favor the death

event in ICU. While the variables the use of corticosteroids and platelets (3rd day) are associated with a longer survival.

For the case of Cause-specific hazard model, the model considers that all the variables are significant. The variables that are associated with a higher risk of death in ICU are: *equal_ICU_IMV_times*, age, lymphocytes (ICU), creatinine (ICU, day3), acute kidney failure, use of statins or interferon beta, neuromuscular blockers, hemodynamic SOFA (ICU), diabetes and flu vaccine. While corticosteroids and flu vaccine are associated with increased survival. In any case, these values are biased since the subjects who experience the competitive event (leaving the ICU due to improvement) are censored at the time of the occurrence of the competing event.

variable	coeff	exp(coeff)	z	p-value
equal_ICU_IMV_times	4.24e+0	69.417	15.341	0.0e+00
age	2.05e-02	1.021	3.243	1.2e-03
hepaticDysfunction	-9.41e-02	0.910	-0.945	3.4e-01
urea_day3	2.23e-03	1.002	1.328	1.8e-01
timeInIMV	-6.83e-02	0.934	-4.403	1.1e-05
glucose_day3	1.78e-03	1.002	2.206	2.7e-02
crp_day3	6.70e-03	1.007	1.530	1.3e-01
lymphocytes_ICU	6.26e-02	1.065	1.998	4.6e-02
crp_ICU	-3.89e-03	0.996	-0.872	3.8e-01
creatinine_day3	4.21e-02	1.043	0.762	4.5e-01
acuteKidneyFailure	4.98e-01	1.645	4.964	6.9e-07
PaFi_ICU	-1.01e-03	0.999	-1.339	1.8e-01
ventilatoryRatio_day3	5.36e-02	1.055	0.917	3.6e-01
corticosteroid	-3.73e-01	0.689	-3.048	2.3e-03
statins	1.86e-01	1.204	1.772	7.6e-01
Interferon_beta	5.24e-03	1.005	0.050	9.6e-01
Tocilizumab	-5.05e-02	0.951	-0.442	6.6e-01
NeuroblockNeeded	4.71e-01	1.601	2.298	2.2e-02
septic_shock_day3	1.32e-02	1.013	0.123	9.0e-01
diabetes	6.08e-02	1.063	0.531	6.0e-01
fluVaccine	-1.80e-01	0.835	-1.300	1.9e-01
Infectious_complications_Lungs	-3.53e-02	0.965	-0.370	7.1e-01
platelets_day3	-2.41e-03	0.998	-3.594	3.3e-04
bacteremia	-7.86e-02	0.924	-0.712	4.8e-01
sofa_hemo_ICU	2.22e-02	1.022	0.821	4.1e-01

Table 7.12: Subdistribution hazard model [1/2]

variable	coeff	exp(coeff)	z	p-value
platelets_ICU	-2.76e-04	1.000	-0.552	5.8e-01
previousRespiratorySupport Mascarilla_de_oxígeno_con _bolsa_reservorio	9.53e-02	1.100	0.994	3.2e-01
sofa_hemo_day3	5.67e-02	1.058	1.834	6.7e-02
leucocytes_day3	-7.13e-03	0.993	-0.644	5.2e-01
Lopinavir_ritonavir	4.91e-02	1.050	0.386	7.0e-01
recruitmentManouvers	-6.00e-02	0.942	-0.552	5.8e-01
creatinine_ICU	-5.10e-02	0.950	-0.641	5.2e-01
glucose_ICU	-2.60e-04	1.000	-0.403	6.9e-01
variablesLdh_ICU	2.66e-05	1.000	0.206	8.4e-01
procalcitonin_ICU	-6.74e-04	0.999	-0.083	9.3e-01
Hospital_Admission Start_Symptoms	-7.67e-03	0.992	-0.775	4.4e-01
Num. cases = 1129 Pseudo Log-likelihood = -2338 Pseudo likelihood ratio test = 1277 on 37 df				

Table 7.13: Subdistribution hazard model [2/2]

variable	coeff	exp(coeff)	z	p-value
equal_ICU_IMV_times	5.629e+00	2.784e+02	18.740	0.0
age	7.763e-03	1.008e+00	1.286	0.0
hepaticDysfunction	-2.602e-02	9.743e-01	-0.228	0.0
urea_day3	1.865e-03	1.002e+00	1.080	0.0
timeInIMV	-1.592e-01	8.528e-01	-18.885	0.0
glucose_day3	8.285e-04	1.001e+00	0.891	0.0
crp_day3	4.950e-03	1.005e+00	1.085	0.0
lymphocytes_ICU	5.247e-02	1.054e+00	1.566	0.0
crp_ICU	-6.432e-04	9.994e-01	-0.136	0.0
creatinine_day3	3.877e-02	1.040e+00	0.605	0.0
acuteKidneyFailure	1.006e-01	1.106e+00	0.842	0.0
PaFi_ICU	-6.754e-05	9.999e-01	-0.098	0.0
ventilatoryRatio_day3	2.659e-03	1.003e+00	0.043	0.0
corticosteroid	-3.636e-01	6.952e-01	-2.842	0.001
statins	2.178e-01	1.243e+00	1.972	0.01
Interferon_beta	6.654e-02	1.069e+00	0.575	0.01
Tocilizumab	-8.951e-02	9.144e-01	-0.783	0.01
NeuroblockNeeded	7.012e-01	2.016e+00	3.339	0.01
septic_shock_day3	-2.706e-02	9.733e-01	-0.209	0.0
diabetes	5.527e-02	1.057e+00	0.395	0.0
fluVaccine	5.664e-02	1.058e+00	0.406	0.0
Infectious_complications_Lungs	-2.278e-01	7.963e-01	-2.032	0.01
platelets_day3	7.165e-05	1.000e+00	0.113	0.01
bacteremia	-3.273e-01	7.209e-01	-2.866	0.001
sofa_hemo_ICU	5.274e-04	1.001e+00	0.018	0.001

Table 7.14: Cause-specific hazard model [1/2]

variable	coeff	exp(coeff)	z	p-value
platelets_ICU	-2.127e-04	9.998e-01	-0.376	0.001
previousRespiratorySupport Mascarilla_de_oxígeno_con _bolsa_reservorio	1.875e-03	1.002e+00	0.017	0.001
sofa_hemo_day3	-2.072e-03	9.979e-01	-0.063	0.001
leucocytes_day3	7.513e-03	1.008e+00	0.743	0.001
Lopinavir_ritonavir	-1.864e-02	9.815e-01	-0.140	0.001
recruitmentManouvers	-1.446e-01	8.653e-01	-1.273	0.001
creatinine_ICU	6.057e-02	1.062e+00	0.662	0.001
glucose_ICU	-3.601e-04	9.996e-01	-0.469	0.001
variablesLdh_ICU	1.665e-04	1.000e+00	1.014	0.001
procalcitonin_ICU	-2.592e-03	9.974e-01	-0.283	0.001
Hospital_Admission Start_Symptoms	-1.629e-03	9.984e-01	-0.177	0.001
Concordance= 0.949 (se = 0.006) Likelihood ratio test= 1509 on 36 df, p=<2e-16 Wald test = 520.9 on 36 df, p=<2e-16 Score (logrank) test = 1231 on 36 df, p=<2e-16				

Table 7.15: Cause-specific hazard model [2/2]

7.5 Other analyses and methods

This section shows the results of survival analyses prior to approaches A and B (7.5.0.3, 7.5.0.4) and variants of the analyses A and B (7.5.0.1, 7.5.0.2), used as sanity checks and as a ablation study.

7.5.0.1 Approaches A and B without certain correlated features

Approaches A and B have been replicated removing highly correlated variables. However, correlations corresponding to the same variable for different events have not been removed. If both features are clinically important, they are not removed either. Then time variables, urea, lactate, septic shock and hemodynamic SOFA have been remove. Figures 10.126, 10.127 show the resulting correlations for approach A and 10.131, 10.132, 10.133 the resulting correlations for approach B. Significant correlations are framed in red.

The results show that removing these variables the performance decreases both for feature selection and for survival analysis. We believe that this may be due to the fact that these variables can be indicators of the patient's evolution when they are taken into account together with other variables. For this reason, the option of including correlations is preferable.

7.5.0.2 Approaches A and B without time and dates features

Approaches A and B have also been replicated by removing variables for which we would have no information in a real situation during the first three days of ICU. That is, *timeInIMV* and *equal_ICU_IMV_times* are removed. The column *timeInHospital* is kept to model the event of in-hospital death in approach A but removed in approach B, while *timeInICU* is kept in approach B but removed in approach A. It is known that these variables are relevant in the prediction of both events hospital death and ICU discharge. In order to prevent a negative impact on performance, we add new variables that had not been included in approaches A and B but that were taken into account in the initial selection feature documented in 4.1.8. The variables that have been added for both approaches are shown in Table 4.1. The performances of the models that achieve higher C-index in feature selection are shown in Table 7.16. C-indexes oscillate between 0.72-0.79 for approach A and 0.67-0.77 for approach B. The brier scores tend to be higher than in the previous approaches, with the maximum value being 0.09 (approach A) and 0.08 (approach B).

The following Table 7.17 shows the models that have obtained the highest accuracy in the classification and in Table 7.18 those that have obtained the

approach	method	C-index	Brier score
approach A	CSF (<i>max_features = sqrt</i>)	0.79	0.09
approach B	RSF (<i>max_features = sqrt</i>)	0.77	0.07

Table 7.16: Feature selection models with higher C-index for approach A and B

approach	method	class	precision	recall	f1-score	accuracy
approach A	SVM lineal kernel	non-survivor	0.82	0.90	0.86	0.82
		survivor	0.81	0.68	0.74	
approach B	logistic regression	non-survivor	0.81	0.88	0.85	0.81
		survivor	0.80	0.70	0.75	

Table 7.17: Linear classifiers with higher accuracy for approaches A and B

highest C-index in the survival analysis. As expected, C-index remains lower than the ones obtained using time variables. For approach A, there was not much difference between models, the values ranged between 0.74-0.75. While for Brier score, the maximum value was reached by CSF (0.11) and the minimum by nonlinear Cox (0.03). For approach B, the returns are much lower than expected. There is not a single model that reaches a C-index of 0.6. These results are not acceptable thus approach B, with this feature selection, would be discarded.

7.5.0.3 Multivariable Cox Proportional Hazards

Before performing approaches A and B, we implemented a multivariable Cox with three models: baseline model (include comorbidities, laboratory and

approach	method	C-index	Brier score
approach A	non-linear Cox	0.75	0.03
approach B	standard CoxPH	0.54	0.07

Table 7.18: Survival analysis models with higher C-index for approaches A and B

ventilation variables for ICU, previous medication and complementary therapies), a second model for 3rd ICU day with laboratory variables and ventilation and third model with complications.

The features shown in 7.12, 7.13 and 7.14 were chosen through the feature selection with the methods described in 4.1.8 in combination with those selected by clinical experts. Previously, some variables were removed in the correlation analysis. Below each model, the Concordance index, the partial AIC and the result of the log-likelihood ratio test and $-\log_2(p)$ of ll-ratio test are also shown.

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	$-\log_2(p)$
previousRespiratorySupport, Mascarilla facial simple (Hudson)	-1.3196	0.2672	1.0069	-3.2930	0.6539	0.0371	1.9230	-1.3106	0.1900	2.3959
hiv	-0.0452	0.9558	1.0201	-2.0444	1.9541	0.1295	7.0577	-0.0443	0.9647	0.0519
asthma	0.3969	1.4872	0.2362	-0.0661	0.8598	0.9361	2.3627	1.6803	0.0929	3.4281
hypertension	0.0367	1.0374	0.1093	-0.1776	0.2510	0.8373	1.2853	0.3358	0.7370	0.4403
heartChronic	0.3365	1.4000	0.1399	0.0623	0.6106	1.0643	1.8416	2.4056	0.0161	5.9526
pulmonarChronic	0.4801	1.6162	0.1429	0.2001	0.7601	1.2215	2.1385	3.3606	0.0008	10.3286
fluVaccine	-0.2493	0.7794	0.1428	-0.5291	0.0305	0.5892	1.0310	-1.7461	0.0808	3.6296
antibiotic	0.0746	1.0775	0.5909	-1.0835	1.2327	0.3384	3.4305	0.1263	0.8995	0.1528
statins	0.0202	1.0204	0.1133	-0.2019	0.2423	0.8172	1.2742	0.1784	0.8584	0.2202
neuromuscular blockers	0.7198	2.0541	0.1908	0.3458	1.0939	1.4131	2.9858	3.7717	0.0002	12.5904
Interferon beta	0.1346	1.1440	0.1058	-0.0729	0.3420	0.9297	1.4078	1.2715	0.2036	2.2964
ventilatoryRatio_ICU	0.0136	1.0137	0.0517	-0.0878	0.1150	0.9160	1.1219	0.2630	0.7926	0.3354
platelets_ICU	-0.1567	0.8550	0.0575	-0.2694	-0.0440	0.7639	0.9569	-2.7256	0.0064	7.2836
PaFi_ICU	-0.0915	0.9126	0.0562	-0.2016	0.0187	0.8174	1.0189	-1.6270	0.1037	3.2689
variablesLdh_ICU	0.1378	1.1477	0.0489	0.0420	0.2336	1.0429	1.2631	2.8193	0.0048	7.6987
dDimer_ICU	-0.0317	0.9688	0.0518	-0.1332	0.0698	0.8753	1.0723	-0.6128	0.5400	0.8889
lymphocytes_ICU	-0.0287	0.9717	0.0685	-0.1630	0.1055	0.8496	1.1113	-0.4193	0.6750	0.5671
septic_shock_ICU	0.0829	1.0864	0.0456	-0.0065	0.1723	0.9936	1.1880	1.8184	0.0690	3.8571
Concordance	0.6271									
Partial AIC	3274.6605									
log-likelihood ratio test	67.1729 on 18 df									
$-\log_2(p)$ of ll-ratio test	22.8192									

Figure 7.12: Baseline model for multivariable Cox Proportional Hazards

Among the statistically significant variables we find heart chronic, pulmonary chronic, the use of neuromuscular blockers and LDH (ICU) stand out as risk factors for the event of hospital death. For the 3rd day at ICU event, LDH and septic shock are associated with a higher risk while paFi is associated with a longer survival. Regarding complications, pulmonary coinfections and hypertension also appear as risk factors.

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
PaFi_day3	-0.2368	0.7892	0.0570	-0.3486	-0.1250	0.7057	0.8825	-4.1515	<5e-05	14.8859
variablesLdh_day3	0.1264	1.1347	0.0462	0.0358	0.2170	1.0364	1.2423	2.7344	0.0062	7.3220
dDimer_day3	-0.0730	0.9296	0.0554	-0.1817	0.0357	0.8339	1.0363	-1.3167	0.1879	2.4116
lymphocytes_day3	0.0295	1.0299	0.0468	-0.0622	0.1211	0.9397	1.1287	0.6301	0.5286	0.9196
septic_shock_day3	0.1203	1.1278	0.0485	0.0253	0.2153	1.0256	1.2402	2.4813	0.0131	6.2553
Concordance	0.6029									
Partial AIC	3220.7147									
log-likelihood ratio test	36.1011 on 5 df									
-log2(p) of ll-ratio test	20.0731									

Figure 7.13: 3r ICU day model for multivariable Cox Proportional Hazards

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
dic	0.3167	1.3726	0.1835	-0.0430	0.6764	0.9579	1.9668	1.7258	0.0844	3.5668
pulmonaryEmbolism	-0.4887	0.6134	0.2100	-0.9003	-0.0770	0.4064	0.9258	-2.3268	0.0200	5.6457
ictus	-0.1272	0.8806	0.2966	-0.7085	0.4541	0.4924	1.5748	-0.4288	0.6680	0.5820
bacteremia	-0.2085	0.8118	0.1073	-0.4189	0.0018	0.6578	1.0018	-1.9431	0.0520	4.2654
Infectious complications: Lungs	-0.3507	0.7042	0.1083	-0.5631	-0.1384	0.5695	0.8707	-3.2375	0.0012	9.6957
hypertension	0.3323	1.3942	0.1020	0.1325	0.5322	1.1417	1.7026	3.2593	0.0011	9.8062
Concordance	0.6067									
Partial AIC	4530.1969									
log-likelihood ratio test	34.3245 on 6 df									
-log2(p) of ll-ratio test	17.3898									

Figure 7.14: Complications model for multivariable Cox Proportional Hazards

The training concordance index of the models are 0.62 (baseline), 0.60 (3r ICU day, complications). Those discriminative powers are not good enough. For this reason these models (taking into account each selected feature set), were discarded.

7.5.0.4 Cox's time varying proportional hazard

This model was developed after the Cox multivariate in order to improve C-index. For this purpose, the dataset was transformed, and two rows were added per patient to indicate ICU and 3rd ICU day variables. Data from antibiotics, antivirals (tocilizumab and interferon beta), corticosteroids, neuromuscular blockers, and positive bacterial tests for lung infections were transformed as well. In Figures 7.15 and 7.16 results for multivariable Cox's time varying proportional hazard are shown.

The variables that appear associated with worse survival are: creatinine, age, chronic heart, chronic pulmonary, use of neuromuscular blockers, cardiac

arrest and platelets, paFi and LDH appear with $HR = 1$ which means that they have no effect on survival. The use of corticosteroids appears as an indicator of good prognosis. Organizing pneumonia and positive tests in lung coinfection also have a HR lower than one, however, these variables are not associated with a positive prognosis. These variables are considered variables with a protective effect (variables in which survival increases). This phenomena can occur when patients die before developing the disease, although in our case this does not happen. In the statistical analysis, we saw that 3.73% of patients develop organizing pneumonia and 41.99% experience pulmonary coinfections. It is therefore concluded that the inference process of the model is not sufficiently precise.

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	- log2(p)
ventilatoryRatio	0.15	1.16	0.06	0.04	0.26	1.04	1.29	2.61	0.01	6.78
platelets	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	-5.13	<0.005	21.70
PaFi	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	-4.39	<0.005	16.44
creatinine	0.24	1.27	0.04	0.16	0.32	1.18	1.38	6.07	<0.005	29.53
variablesLdh	0.00	1.00	0.00	0.00	0.00	1.00	1.00	3.89	<0.005	13.30
dDimer	-0.00	1.00	0.00	-0.00	0.00	1.00	1.00	-0.60	0.55	0.87
lymphocytes	0.03	1.03	0.04	-0.05	0.10	0.95	1.11	0.74	0.46	1.12
septic_shock	0.15	1.16	0.12	-0.08	0.38	0.92	1.47	1.28	0.20	2.32
antibiotic	-0.21	0.81	0.13	-0.46	0.03	0.63	1.03	-1.71	0.09	3.51
age	0.04	1.04	0.01	0.03	0.05	1.03	1.05	6.81	<0.005	36.60
previousRespiratorySupport, Mascarilla facial simple (Hudson)	-0.50	0.60	1.00	-2.45	1.45	0.09	4.25	-0.51	0.61	0.71
hiv	-0.04	0.96	0.71	-1.44	1.36	0.24	3.89	-0.06	0.95	0.07
asthma	0.31	1.37	0.22	-0.12	0.74	0.89	2.11	1.42	0.16	2.68
hypertension	0.05	1.05	0.11	-0.16	0.25	0.85	1.29	0.44	0.66	0.59
heartChronic	0.26	1.29	0.13	-0.00	0.52	1.00	1.68	1.93	0.05	4.22
pulmonarChronic	0.40	1.49	0.14	0.13	0.66	1.14	1.94	2.95	<0.005	8.27
fluVaccine	-0.14	0.87	0.14	-0.40	0.13	0.67	1.14	-1.01	0.31	1.67

Figure 7.15: Multivariable Cox's time varying proportional hazard [1/2]

statins	-0.07	0.93	0.11	-0.28	0.14	0.75	1.15	-0.68	0.49	1.02
corticosteroid	-0.33	0.72	0.10	-0.53	-0.12	0.59	0.88	-3.13	<0.005	9.16
neuromuscular blockers	0.61	1.83	0.18	0.25	0.96	1.29	2.61	3.36	<0.005	10.30
Tocilizumab	0.05	1.05	0.15	-0.25	0.34	0.78	1.41	0.31	0.75	0.41
Interferon beta	0.36	1.43	0.25	-0.13	0.85	0.88	2.33	1.43	0.15	2.70
cardiacArrest	1.00	2.73	0.14	0.74	1.27	2.09	3.56	7.39	<0.005	42.60
dic	0.18	1.20	0.18	-0.17	0.54	0.84	1.72	1.01	0.31	1.69
pulmonaryEmbolism	-0.40	0.67	0.21	-0.81	0.01	0.45	1.01	-1.91	0.06	4.16
ictus	-0.03	0.97	0.29	-0.60	0.53	0.55	1.71	-0.12	0.91	0.14
organizingPneumonia	-0.56	0.57	0.26	-1.08	-0.05	0.34	0.95	-2.15	0.03	4.97
Infectious complications: Lungs	-0.24	0.79	0.13	-0.50	0.02	0.61	1.02	-1.78	0.08	3.73
bacteremia	-0.16	0.86	0.11	-0.38	0.07	0.69	1.07	-1.38	0.17	2.58
bacteria_positive_test_with_lung_infection	-0.38	0.68	0.15	-0.69	-0.08	0.50	0.92	-2.49	0.01	6.28

Partial AIC	5337.62
log-likelihood ratio test	359.14 on 30 df
-log2(p) of ll-ratio test	190.46

Figure 7.16: Multivariable Cox's time varying proportional hazard [2/2]

7.6 Summary

This section has been included as a summary to show the performances at metric and selected features level of all survival models implemented within this project. Performance of survival models has been evaluated using two metrics: C-index and Brier score 6.0.2.

Multivariable Cox Proportional Hazards 7.5.0.3 and Cox time varying Proportional Hazards 7.5.0.4 were the first models that we implemented. Among the statistically significant variables, Multivariable Cox Proportional Hazards showed that heart chronic, pulmonary chronic, the use of neuromuscular blockers and LDH (ICU) stand out as risk factors for the event of hospital death. For the 3rd day at ICU event, LDH and septic shock are associated with a higher risk while paFi is associated with a longer survival. Regarding complications, pulmonary coinfections and hypertension also appear as risk factors. Multivariable Cox Proportional Hazards was developed over three models for which we obtained the following C-indexes: 0.62 (baseline model), 0.60 (3r ICU day, complications models).

For Cox time varying Proportional Hazards, the variables that appear associated with worse survival are: creatinine, age, chronic heart, chronic pulmonary, use of neuromuscular blockers, cardiac arrest and platelets, paFi and LDH appear with $HR = 1$ which means that they have no effect on survival. The use of corticosteroids appears as an indicator of good prognosis. Organizing pneumonia and positive tests in lung coinfection also have a HR lower than one,

however, these variables are not associated with a positive prognosis. These variables are considered variables with a protective effect (variables in which survival increases). In the statistical analysis, we saw that 3.73% of patients develop organizing pneumonia and 41.99% experience pulmonary coinfections. It is therefore concluded that the inference process of the model is not sufficiently precise.

Approaches A and B 6.0.3 were subsequently implemented motivated by the low performances of Multivariable Cox Proportional Hazards and Cox time varying Proportional Hazards. We believe that these results were caused firstly by using an inaccurate feature selection and secondly by having too high a percentage of censored data (*i.e.*, survivors). For this reason we decided to use other feature selection such as survival forests: Conditional Survival Forest, Random Survival Forest and Extra-Randomized Survival Forest were used for the selection of new variables for both approaches. CSF (0.90 C-index, 0.03 bs) and RSF(0.85 C-index, 0.04 bs) were the models that achieved the highest C-index for approach A and B respectively.

Regarding the linear classifiers, we observed that the best performances were achieved by SVM linear (0.94 accuracy approach A, 0.97 accuracy approach B). Finally, CSF (0.89 C-index, 0.05 bs) and non-linear Cox (0.86 C-index, 0.04) stood out as the best survival models.

Regarding risk factors, approach A found that CRP (ICU, 3rd ICU day), SOFA score (3rd ICU day), presence of symptoms and lymphocytes (ICU) are associated with a higher risk of death, and platelets (ICU), glucose (ICU), lactate (ICU) to a lesser extent. As variables associated with a good prognosis we find *timeInICU*, *timeInIMV*, paFi (day3) and leukocytes (day3). Longer times of hospitalization or IMV does not have a negative impact on life expectancy and as paFi increases, it can be considered that the quality of the patient's breathing improves.

Risk factors for approach B were analyzed by using models for competitive events, suitable for the treatment of data with competitive events (such as leaving the ICU due to improvement). It considered that *equal_ICU_IMV_times*, age, glucose (3rd day), lymphocytes (ICU), acute kidney failure, neuromuscular blockers requirement, hemodynamic SOFA (3rd day) as risk factors. While the variables the use of corticosteroids and platelets (3rd day) were associated with a longer survival.

At the end, approaches A and B were replicated removing highly correlated variables 7.5.0.1 and time data variables 7.5.0.2. For the first case, we observed that performance decreased both for feature selection and for survival analysis. We believe that this may be due to the fact that these variables can be indicators of the patient's evolution when they are taken into account together

with other variables. For this reason, the option of including correlations was preferable. Removing data time variables performance decreased as well. C-indexes oscillated between 0.72-0.79 for approach A and 0.67-0.77 for approach B. The brier scores tended to be higher than in the previous approaches, with the maximum value being 0.09 (approach A) and 0.08 (approach B).

Chapter 8

Conclusions

This thesis has first performed a retrospective multicenter analysis to characterize patients admitted to the ICU during the first wave of COVID-19 (considering from February 1 to July 31), analyze the risk factors involved in hospital mortality, and develop methods to predict it using statistical analysis and survival methods.

The present study is based on the work carried out in the context of CIBERES-UCI-COVID. CIBERES-UCI-COVID project was awarded in may, 2020, funded by ISCIII. This project gathers data from 69 different Spanish ICUs, including several specific sources such as Getafe hospitals and the SEMICYUC consortium thus becoming the largest data collection effort for ICU data in Spain. In this context, we were also interested in developing a complete and unified database to store data from those hospitals and perform pre-processing analysis comprising the analysis of missing values, outliers, correlations and feature selection.

For the study we selected those patients who required invasive mechanical ventilation during the first day of admission to the ICU and who remained ventilated 3 days later. This interest is motivated by the absence of published studies analyzing the influence of laboratory and ventilatory variables on the third ICU day. In addition, information about baseline (*i.e.*, symptoms, comorbidities, previous medication, etc), outcome and gender must be available. This information is required in order to avoid completely unfilled patients. In retrospect, this decision was not enough to ensure the completeness of the data; a review and correction by the *data entries* of each hospital was necessary to complete information regarding the dates of admission to the ICU, IMV, etc.

Before performing the statistical analysis and survival analysis, the data went through a filtering process that allowed us to define the population of

interest. The final population consisted of 1,140 patients. This amount of patients represents a considerable and acceptable number for conducting scientific research. The size of a sample influences the precision of our estimates and the power of the study to draw conclusions. Using a small sample size may have a negative impact over both of these aspects.

Afterwards, outliers were omitted and then imputed together with the other missing data. Multiple Imputation by Chained Equations (MICE) was used as imputation technique obtaining successful results. At the end of pre-processing, a correlation analysis and a feature selection analysis were performed to remove correlated variables that could impact the performance of machine learning models. However, some of our experiments show that the elimination of correlated variables had a negative impact on the subsequent performance of the survival models. Among the methods for feature selection, traditional methods (logistic regression, random forest and recursive feature elimination) were initially used, but finally survival forests were chosen. This decision was motivated by the results obtained from the first survival models. We hypothesize that the selection of features was not adequate since we obtained successful results with a different set of features while maintaining the same survival models, in this case StandardCox.

The statistical analysis was divided into 7 parts according to the group of variables to be analyzed. Missing values for each feature are ignored. Of the 1,140 patients we observed that compared with survivors, non-survivors were more likely to be older and male (10.40).

Hypertension appears as the most common comorbidity, followed by obesity and diabetes (10.41). Among the most common symptoms we find fever, dry cough, shortness of breath, fatigue and muscle pain (10.42). Important differences between survivors and non-survivors are observed for chronic kidney disease, heart chronic disease and pulmonary chronic disease (10.41). These values are in agreement with those reported by other studies detailed in 5.

Laboratory findings suggest that leukocytes count, CRP, LDH, d-dimer, NT-proBNP, urea are higher in non-survivors suggesting more severe systemic inflammation, cell injury, coagulopathy, risk of cardiac failure and uremia. Regarding the ventilation variables on the first day of mechanical ventilation, no differentiating facts were observed between survivors and non-survivors. However, at the end of ventilation a trend of improvement is observed in survivors and a deterioration for non-survivors. This is reflected as an increase showing signs of respiratory failure (*i.e.*, increase in paused respiratory rate, driving pressure, plateau pressure, paCO_2) and a decrease in variables related to oxygen level and correct physiology of lungs (*i.e.*, paFi , oxygen saturation, compliance) for non-survivors.

Comparing these results with those of studies performed in Europe and China, we come to the conclusion that most common comorbidities, symptoms and treatments remain the same between countries but there are differences in the clinical biomarkers, the durations of IMV, NIMV, and the resulting mortality rates.

Part of the statistical analysis results are reflected in the survival analysis. Two ways of proceeding have been employed, on the one hand we use the event of hospital death and length of hospital stay (procedure named approach A) and on the other hand ICU discharge and length of ICU stay as event and time columns respectively (named approach B).

In order to analyze both events, we decided to employ several survival models: Conditional Survival Forest, Neural SVM, Cox's proportional hazard regression and non-linear Cox. Due to the presence of competitive events in approach B, we decided to use models that take them into account: Cumulative Incidence Function (CIF), Cause-specific hazard model and Fine-Gray Subdistribution hazard model. Linear classifiers have been used to classify patients according to death or improvement. SVM, logistic regression and perceptron have been used and have been validated using 10-fold cross validation.

From a metrics perspective, survival models with the highest C-index correspond to Conditional Survival Forest (*num_trees=200*, *max_features=all*) for approach A and non-linear Cox for approach B obtaining a C-index of 0.89 and 0.86 respectively. Regarding the linear classifiers we find that linear SVM appears as the model that achieves higher values for accuracy and f1-score. The accuracies obtained are 0.96 for approach A and 0.97 for approach B. For approach A, higher accuracies translate into better prediction of survival curves while for approach B it corresponds to the percentage of patients for which the exact day of death is correctly predicted,

In terms of feature relevance with regards to death, StandardCox in approach A (C-index=0.84) shows that CRP (ICU, 3rd ICU day), SOFA score (3rd ICU day), presence of symptoms and lymphocytes (ICU) are associated with a higher risk of death, and platelets (ICU), glucose (ICU), lactate (ICU) to a lesser extent. For approach B, subdistribution hazard model shows as risk factors: age, glucose (3rd day), lymphocytes (ICU), acute kidney failure, neuromuscular blockers requirement and hemodynamic SOFA (3rd day). In addition, the risk of dying is higher for patients who spent the same time in ICU and IMV (*equal_ICU_IMV_times*). As variables associated with a good prognosis we find time spent in ICU, time spent in IMV, paFi (day3) and leukocytes (day3) for StandardCox and the use of corticosteroids and platelets (3rd day) for subdistribution hazard model.

CIF demonstrate that after approximately 50 days, almost all patients in

non-survivors cohort have died. While for the survivors, they tend to spend more time in the ICU. After 100 days almost all surviving patients have been discharged. This aspect is also observed in approach A and earlier in the statistical analysis. Specifically, for approach A, for each unit increase in ICU and IMV time, the risk of death decreases by 71% and 48%, respectively. For the competing risk models, we find that with respect to IMV time, the risk decreases by 6%.

Prior to the present survival analysis, other survival studies were also developed with the Cox Proportional Hazards model (or StandardCox as mentioned above) and Cox's time varying Proportional Hazard model; however, these models did not achieve the expected results and were therefore discarded. We maintain that the causes that induced poor performance were too high percentage of censored patients due to the presence of competing risks (*i.e.*, survivors) and an inaccurate selection of features. The reasons why traditional models do not work well when the number of censored is high due to competitive events is explained in 6.0.1.9.

Subsequent survival studies include variants of approaches A and B taking into account a different selection of features, without taking into account time or length of stay variables and without highly correlated variables. The fact of removing time variables does not affect the linear classifiers too much in practice (accuracies of 0.82 for approach A and 0.81 for approach B) but it does affect survival models. This is due to the strong relevance that time variables have in predicting the death event and the ICU discharge event.

Current results contribute to a better understanding of the behavior of the disease and may guide the implementation of public health measures aiming to improve the management of patients at risk and thus limiting the impact of this pandemic on vulnerable populations. Relevant variables in the prediction of mortality are easily obtainable through blood analysis and PCR tests, thus the reported survival models could be used on a non-profit basis by the vast majority of hospitals internationally.

Chapter 9

Future work

Two lines of action are envisaged for future work. On the one hand, we would like to perform a deeper understanding of current models. Here, the application of explanatory techniques to survival models should be taken into account. An exploration of the state of the art should be carried out and applied interpretability techniques if it is possible. It should also include an analysis of variables that appear as risk factors or factors associated with a better prognosis which indeed are not consistent with real situation. For example, it is the case of leukocytes (3rd ICU day) that appears with a HR above 1 for StandardCox in approach A. This is an indication that a higher number of leukocytes at 3rd ICU day carries lower mortality risk. However, this phenomena is not actually observed in real situation, and statistical analysis showed that the number of leukocytes for non-survivors is higher 5. Although leukocytes always show a non-linear behavior, u-shaped to be exact, these values are not commonly seen. For this reason, one of the future goals is to build a restricted cubic spline that will allow us to model the relationship between leukocytes and the outcome variable.

On the other hand, we are interested in extending both approaches. We would like to combine linear classifiers to competitive risk models and to explore the idea of using ensemble methods to combines all approaches into one.

Chapter 10

Appendix

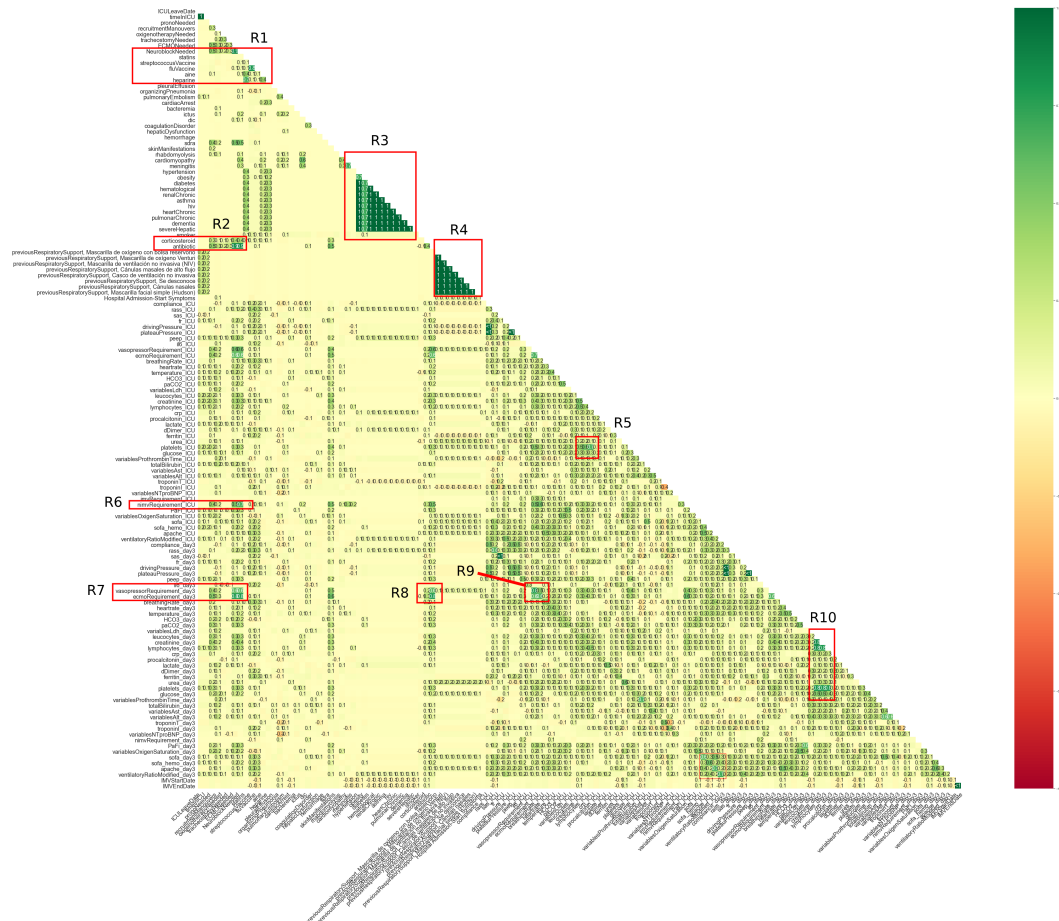


Figure 10.1: Missing correlation plot with captures where the highest correlations are observed

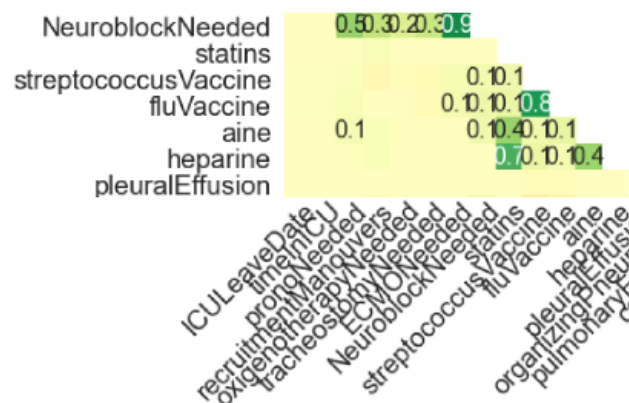


Figure 10.2: Capture R1 of the missing correlation plot

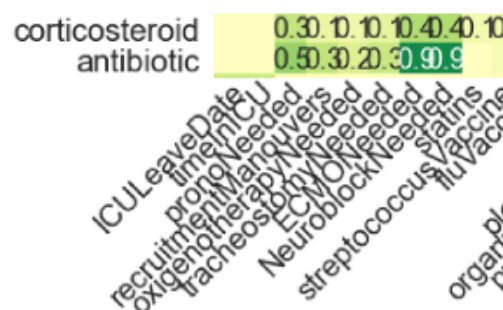


Figure 10.3: Capture R2 of the missing correlation plot

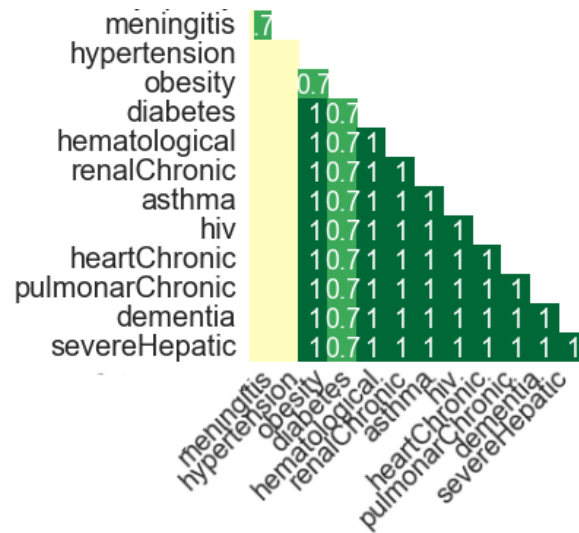


Figure 10.4: Capture R3 of the missing correlation plot

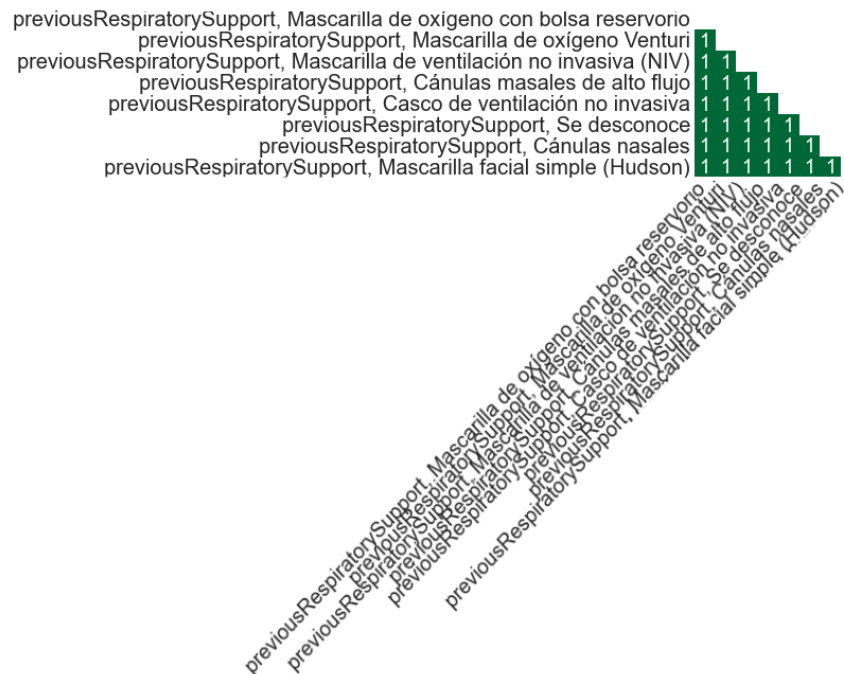


Figure 10.5: Capture R4 of the missing correlation plot

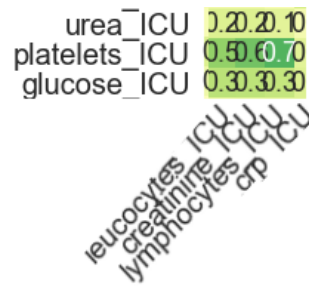


Figure 10.6: Capture R5 of the missing correlation plot

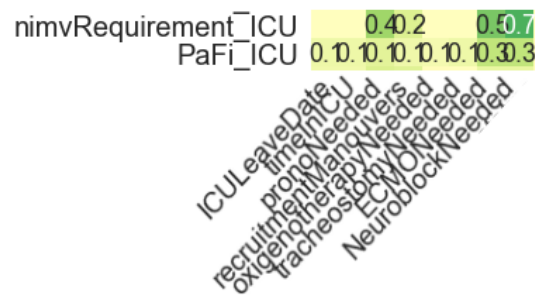


Figure 10.7: Capture R6 of the missing correlation plot

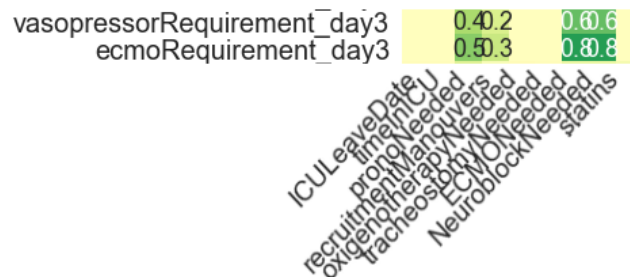


Figure 10.8: Capture R7 of the missing correlation plot



Figure 10.9: Capture R8 of the missing correlation plot

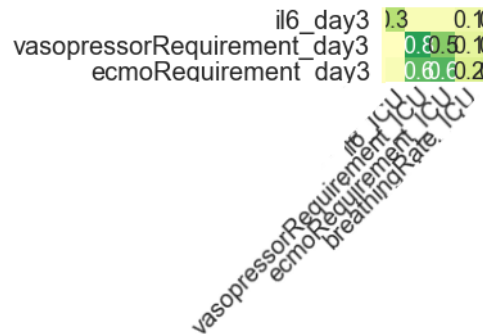


Figure 10.10: Capture R9 of the missing correlation plot

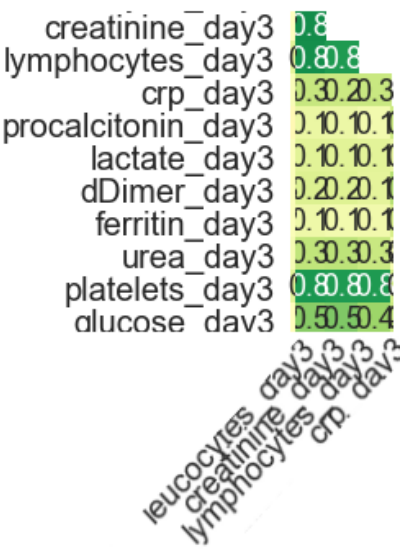


Figure 10.11: Capture R10 of the missing correlation plot

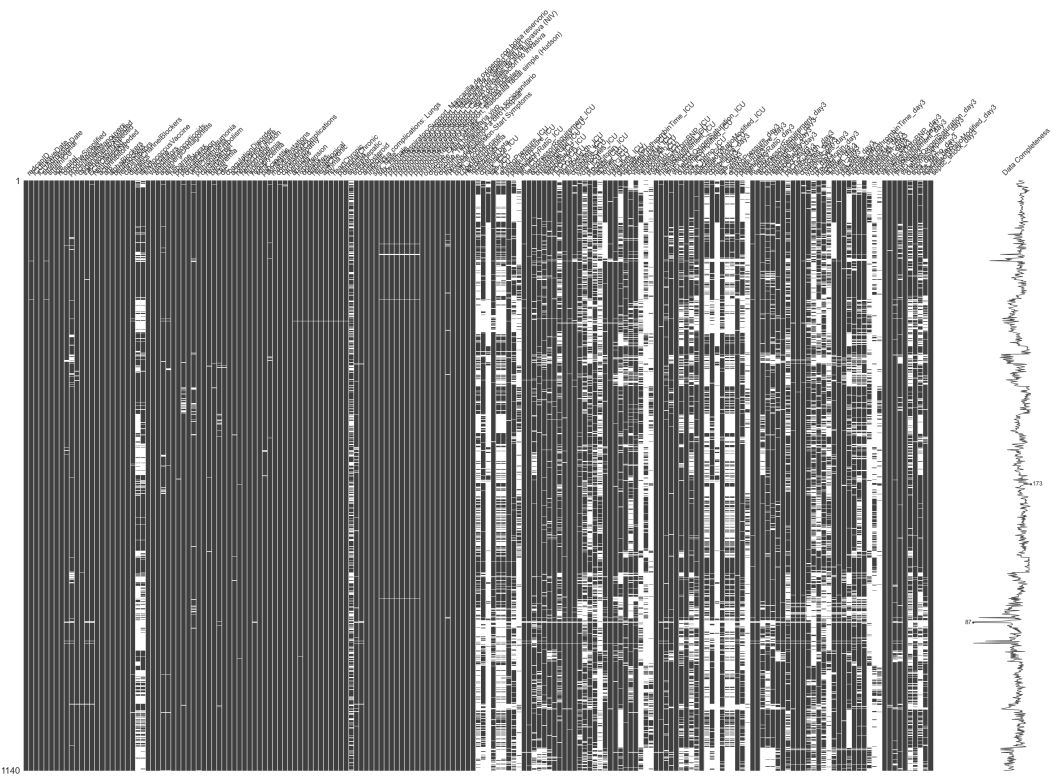


Figure 10.12: Missing values map

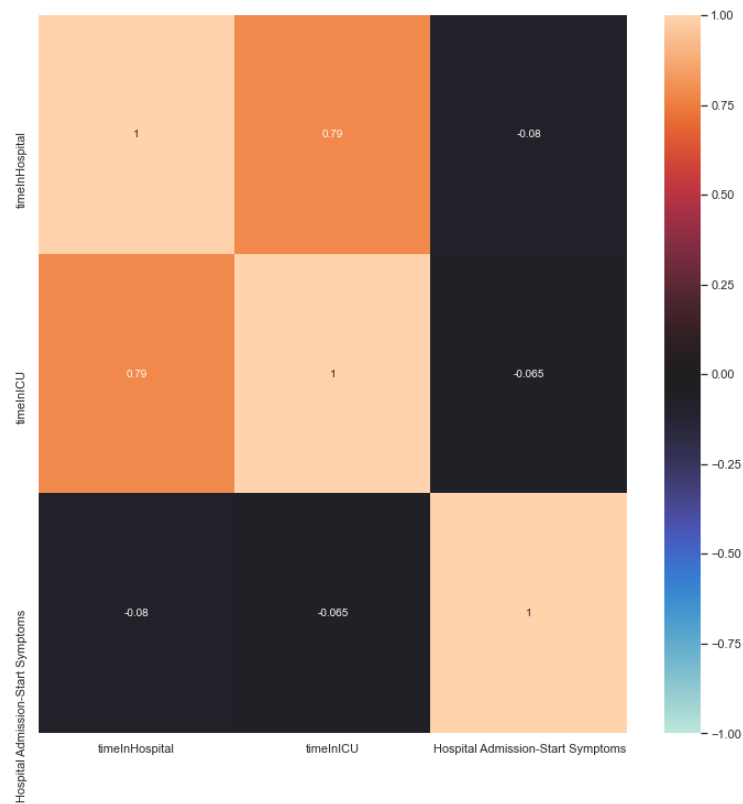


Figure 10.13: Correlation plot of dates and times based variables

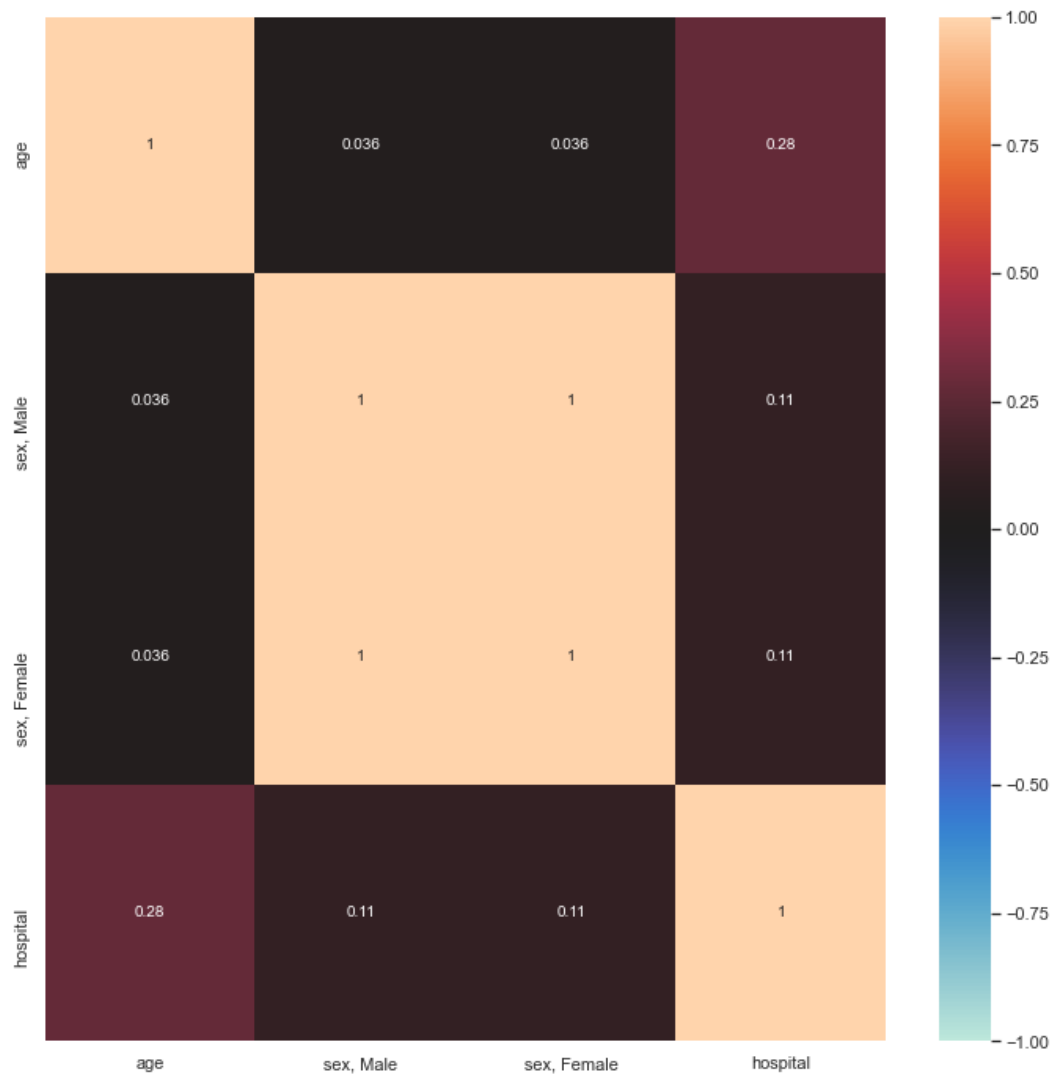


Figure 10.14: Correlation plot for demographic variables

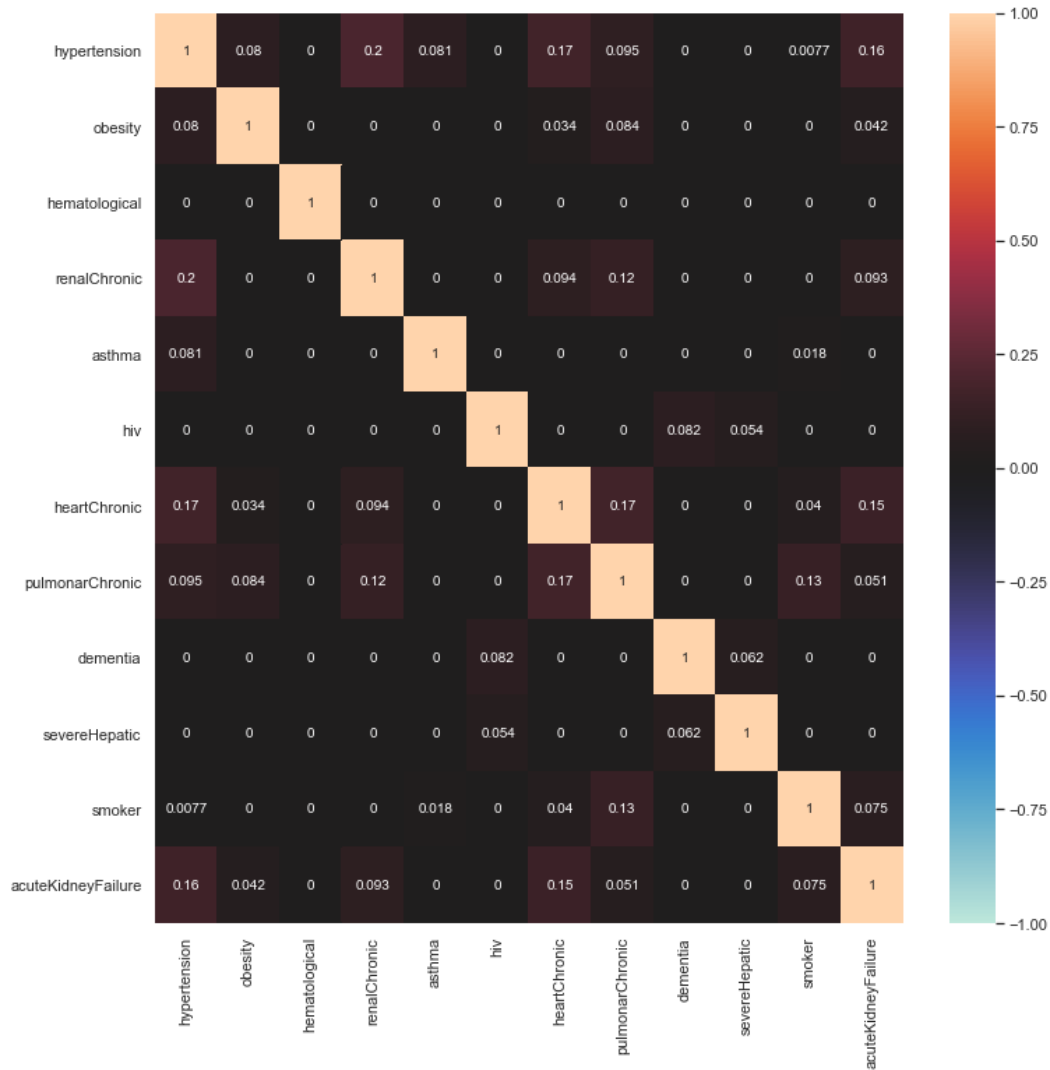


Figure 10.15: Correlation plot for comorbidity variables

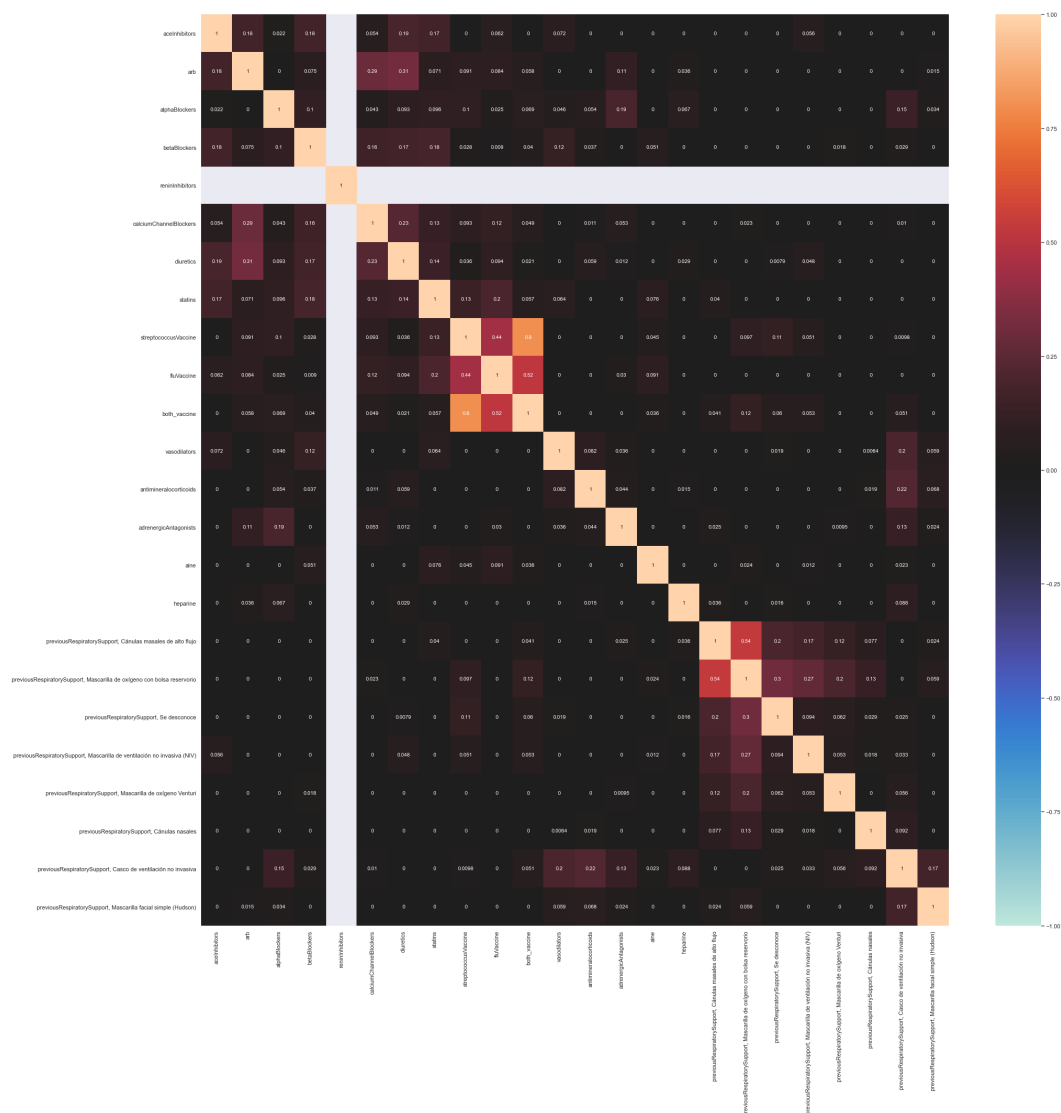


Figure 10.16: Correlation plot for previous medication variables

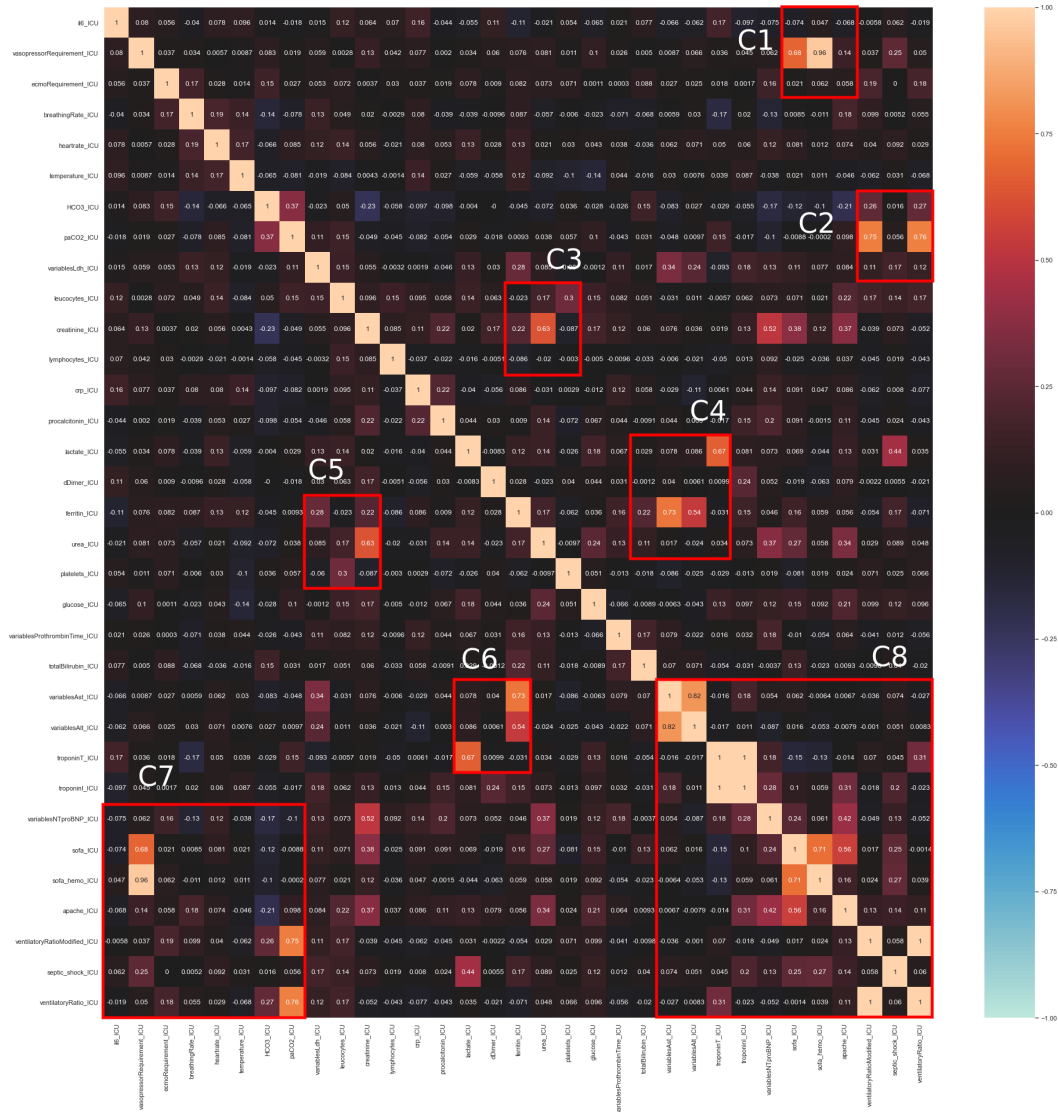


Figure 10.17: Correlation plot for laboratory variables in the first day of ICU

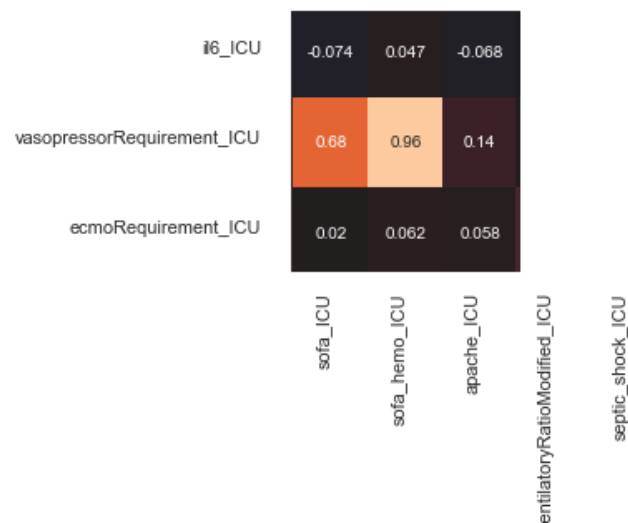


Figure 10.18: Capture C1 of laboratory variables for ICU correlation plot

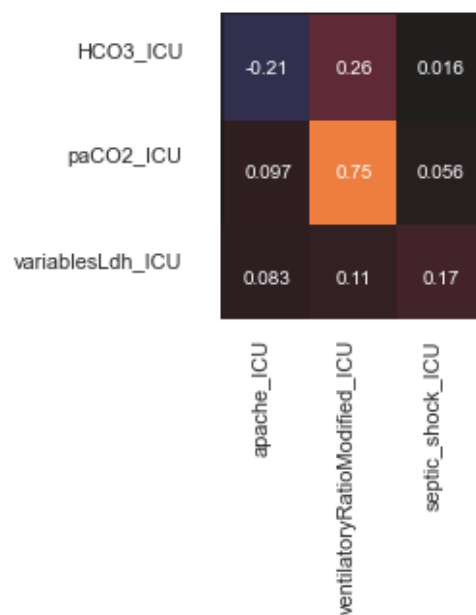


Figure 10.19: Capture C2 of laboratory variables for ICU correlation plot

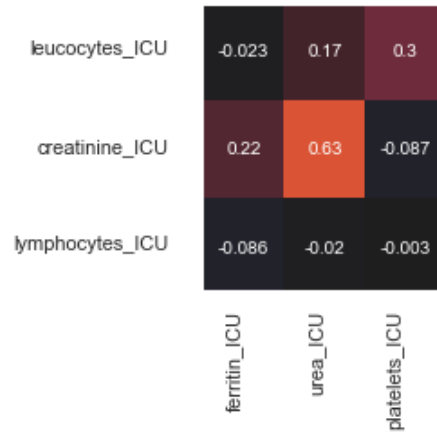


Figure 10.20: Capture C3 of laboratory variables for ICU correlation plot

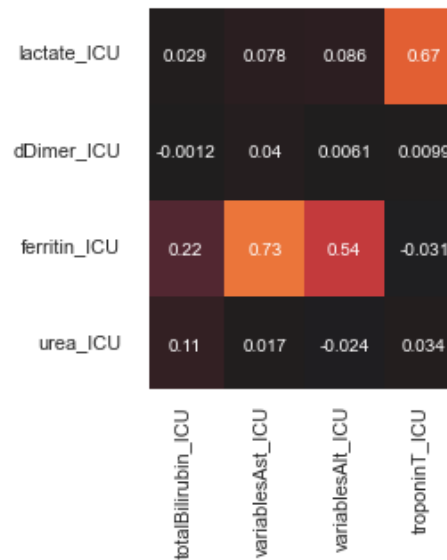


Figure 10.21: Capture C4 of laboratory variables for ICU correlation plot

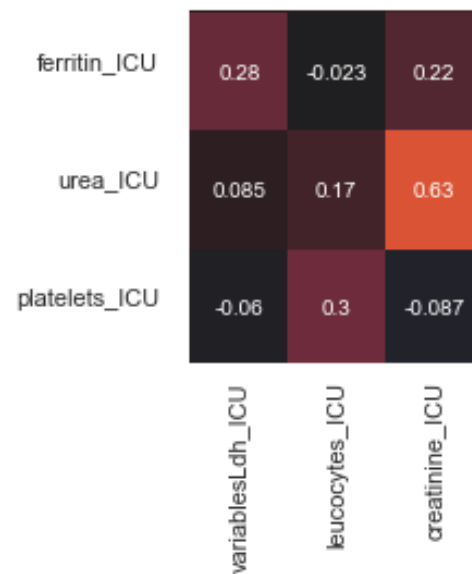


Figure 10.22: Capture C5 of laboratory variables for ICU correlation plot

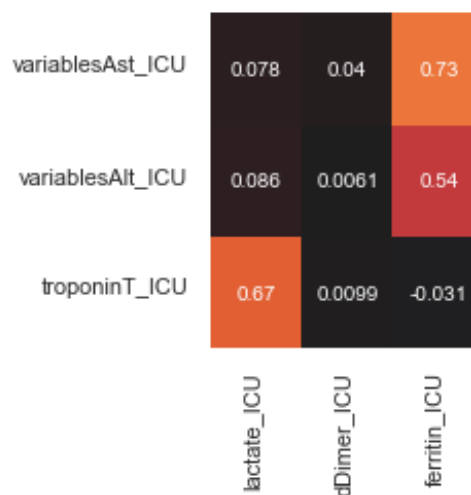


Figure 10.23: Capture C6 of laboratory variables for ICU correlation plot

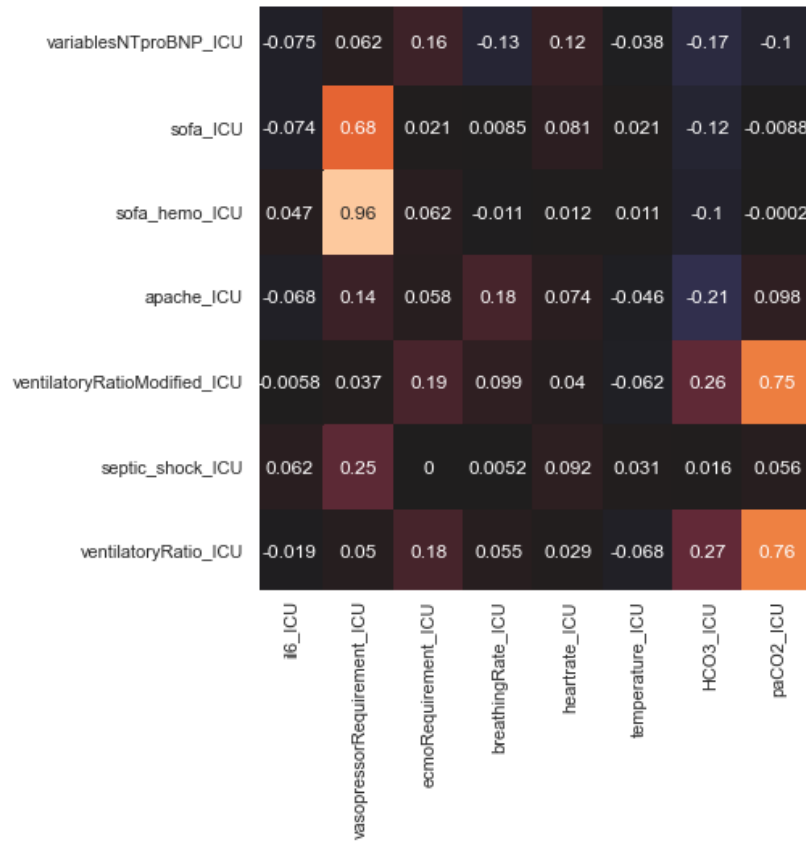


Figure 10.24: Capture C7 of laboratory variables for ICU correlation plot

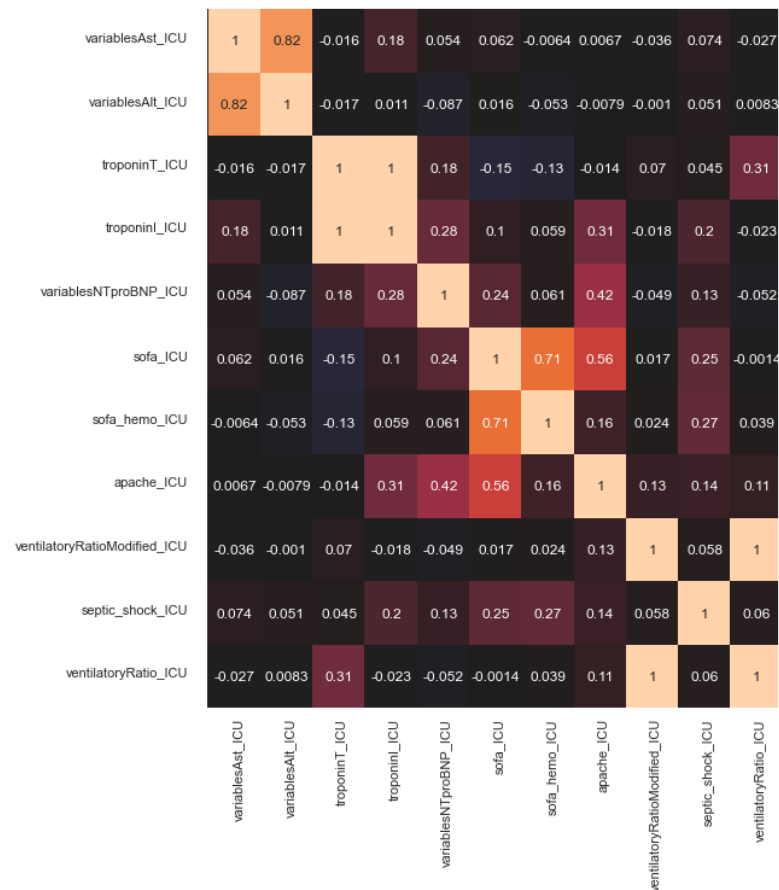


Figure 10.25: Capture C8 of laboratory variables for ICU correlation plot



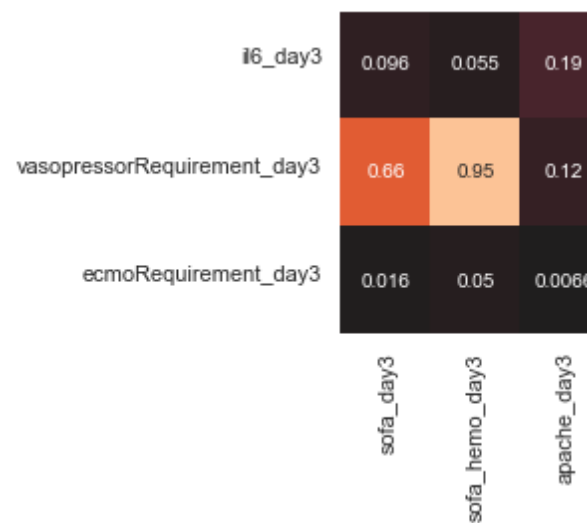


Figure 10.27: Capture C1 of laboratory variables for 3rd correlation plot

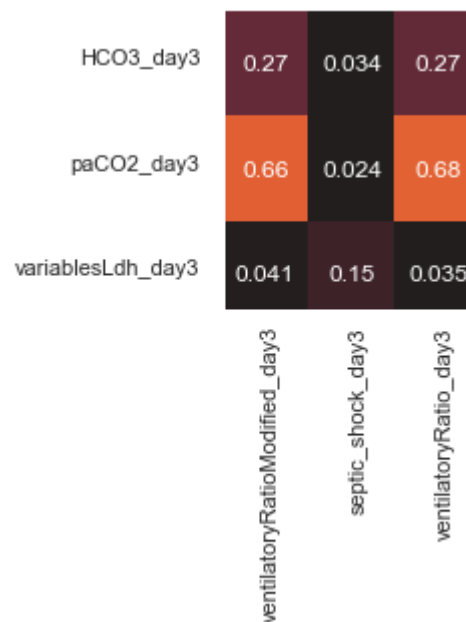


Figure 10.28: Capture C2 of laboratory variables for 3rd correlation plot

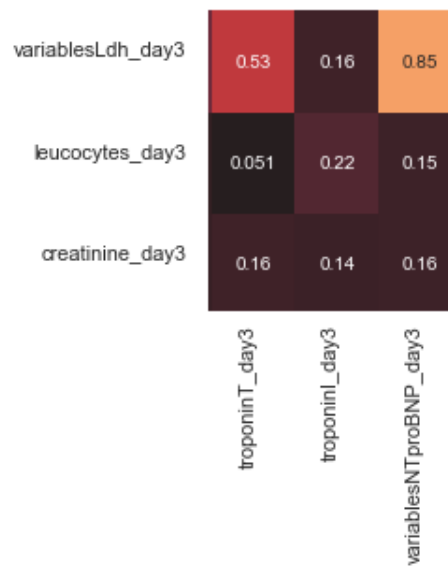


Figure 10.29: Capture C3 of laboratory variables for 3rd correlation plot

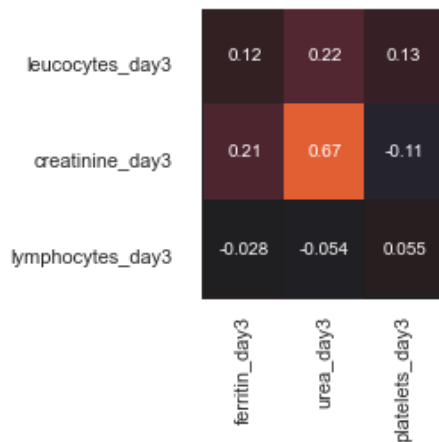


Figure 10.30: Capture C4 of laboratory variables for 3rd correlation plot

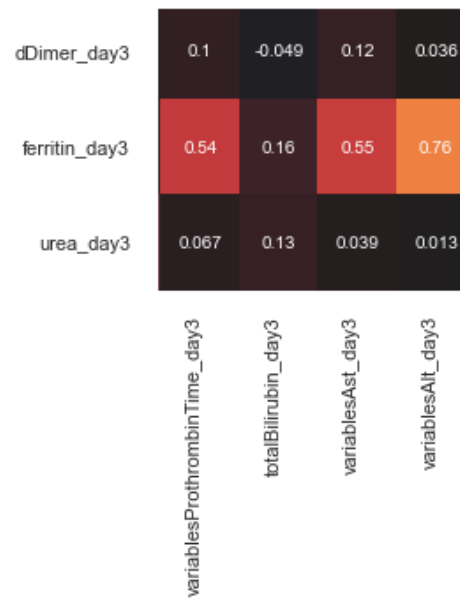


Figure 10.31: Capture C5 of laboratory variables for 3rd correlation plot

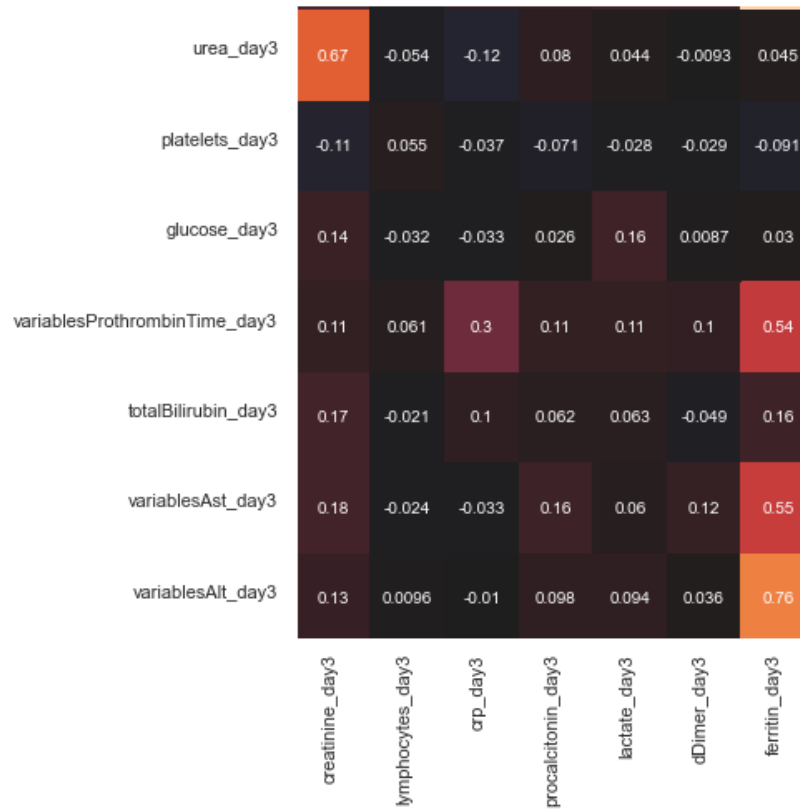


Figure 10.32: Capture C6 of laboratory variables for 3rd correlation plot

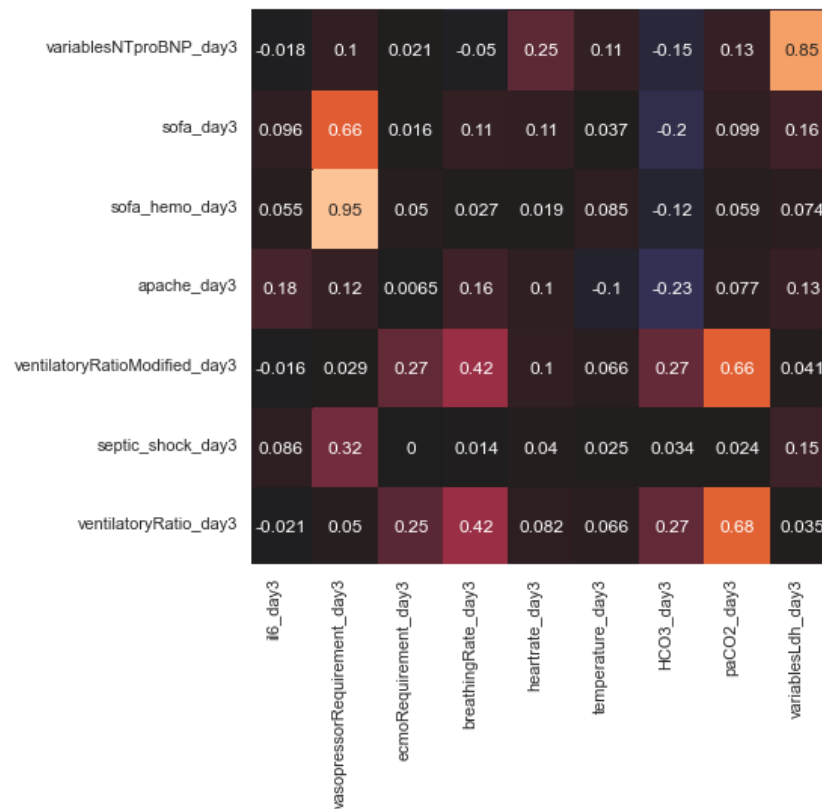


Figure 10.33: Capture C7 of laboratory variables for 3rd correlation plot

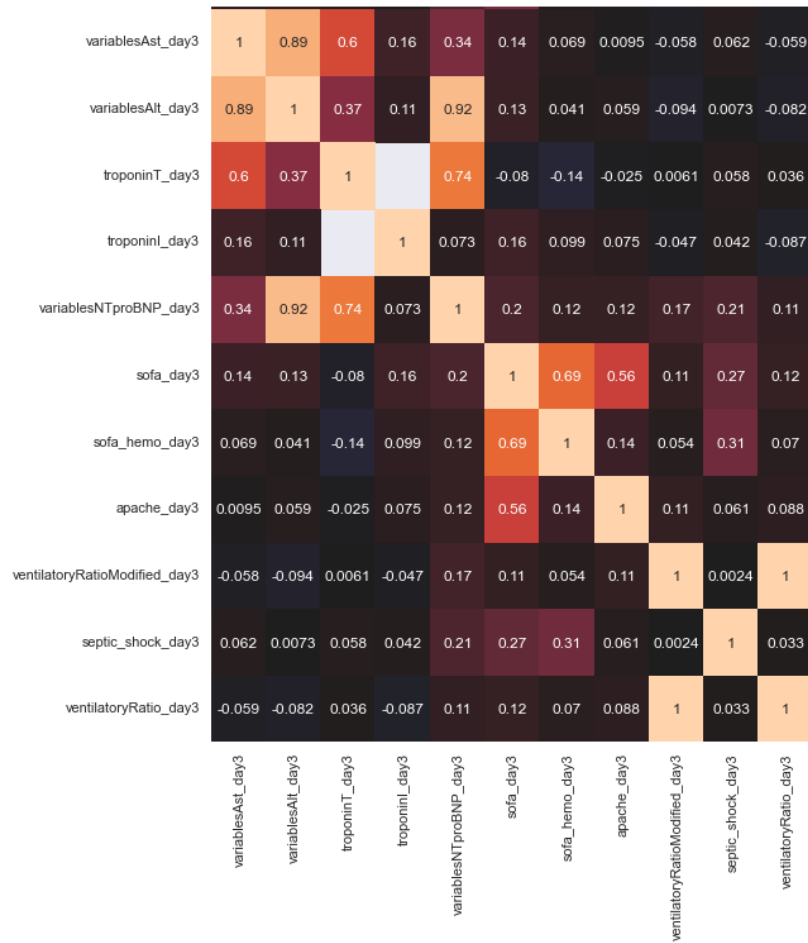


Figure 10.34: Capture C8 of laboratory variables for 3rd correlation plot

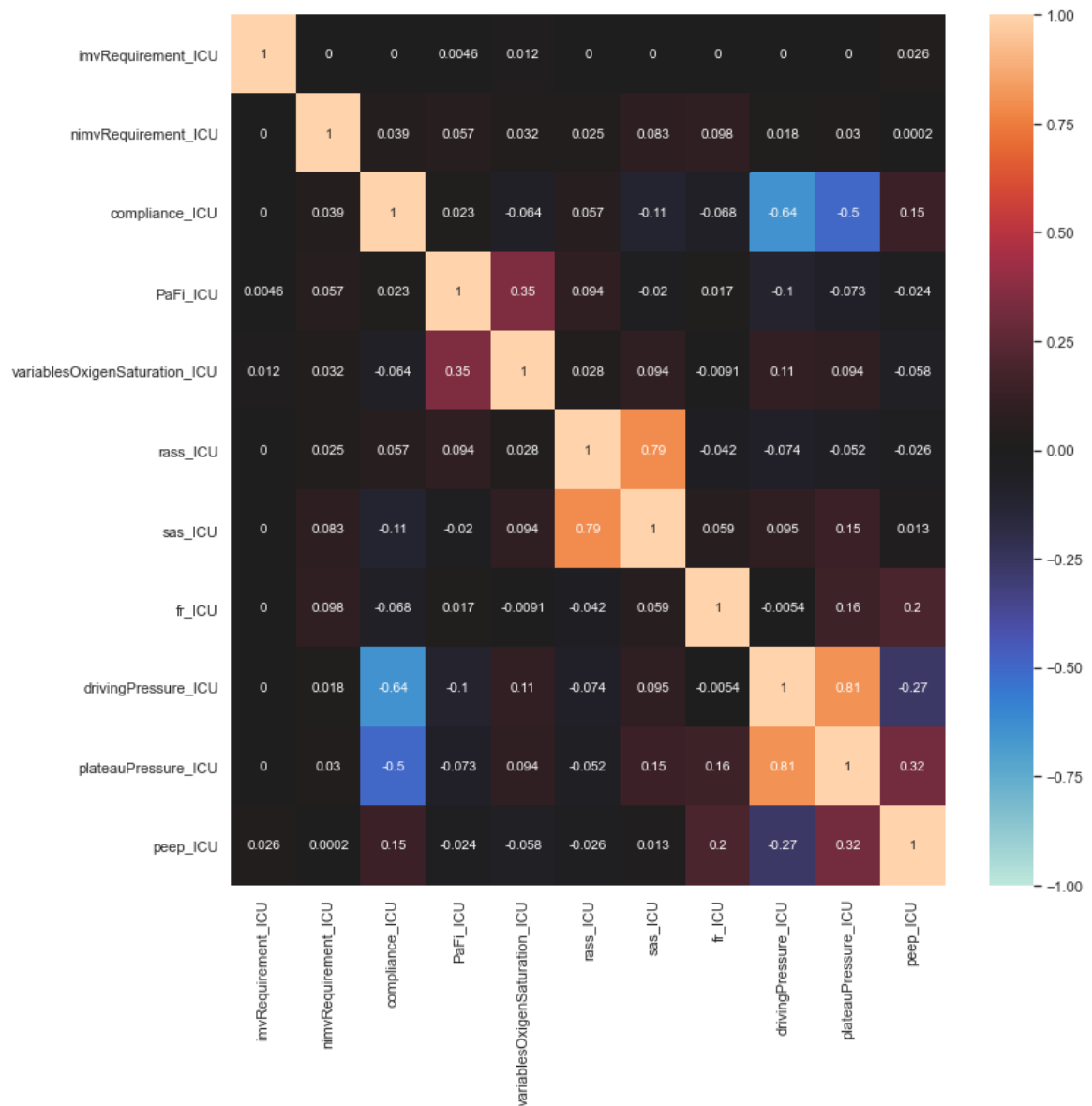


Figure 10.35: Correlation plot for mechanical ventilation variables in the first day of ICU

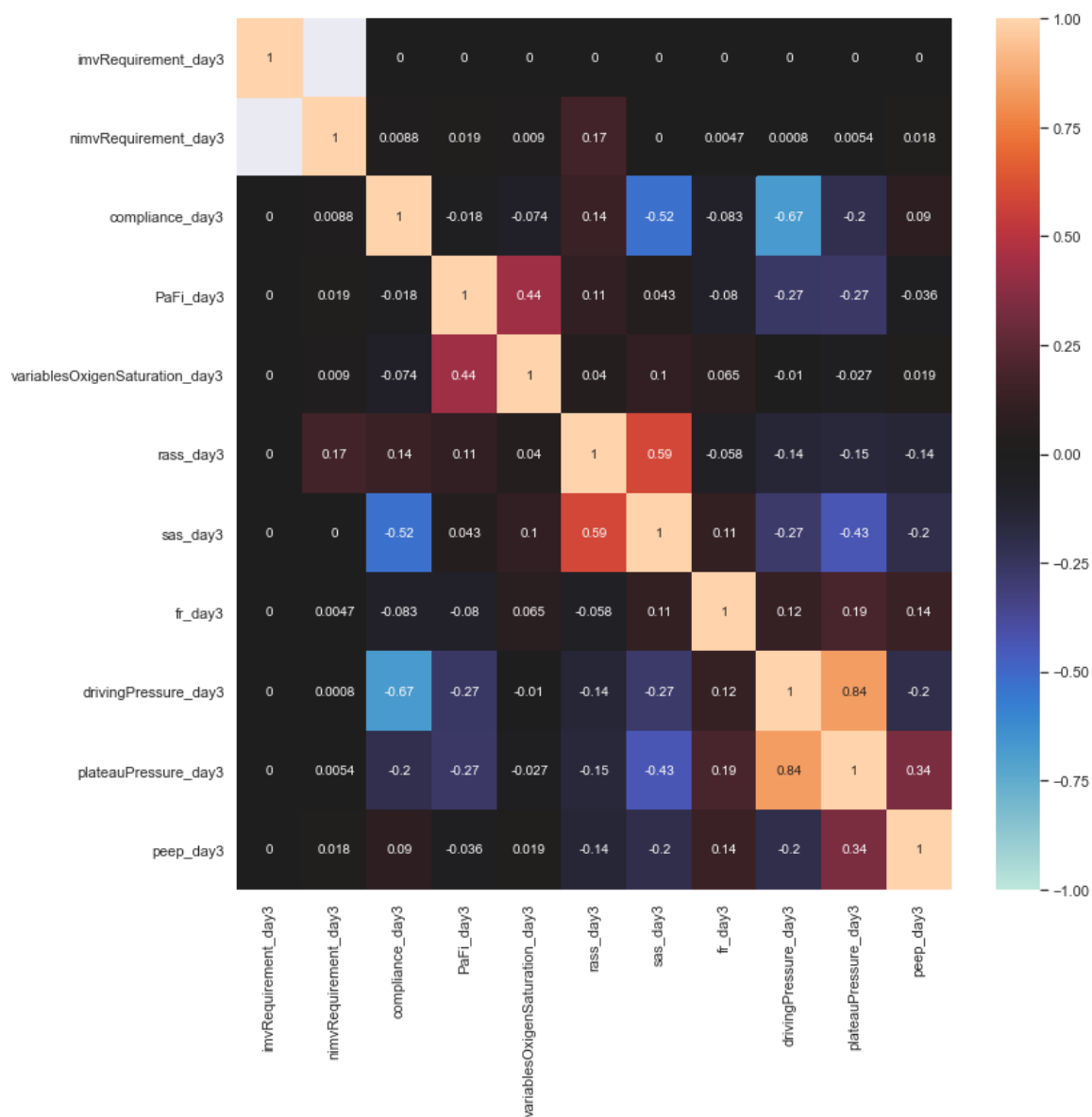


Figure 10.36: Correlation plot for mechanical ventilation variables in the third day of ICU

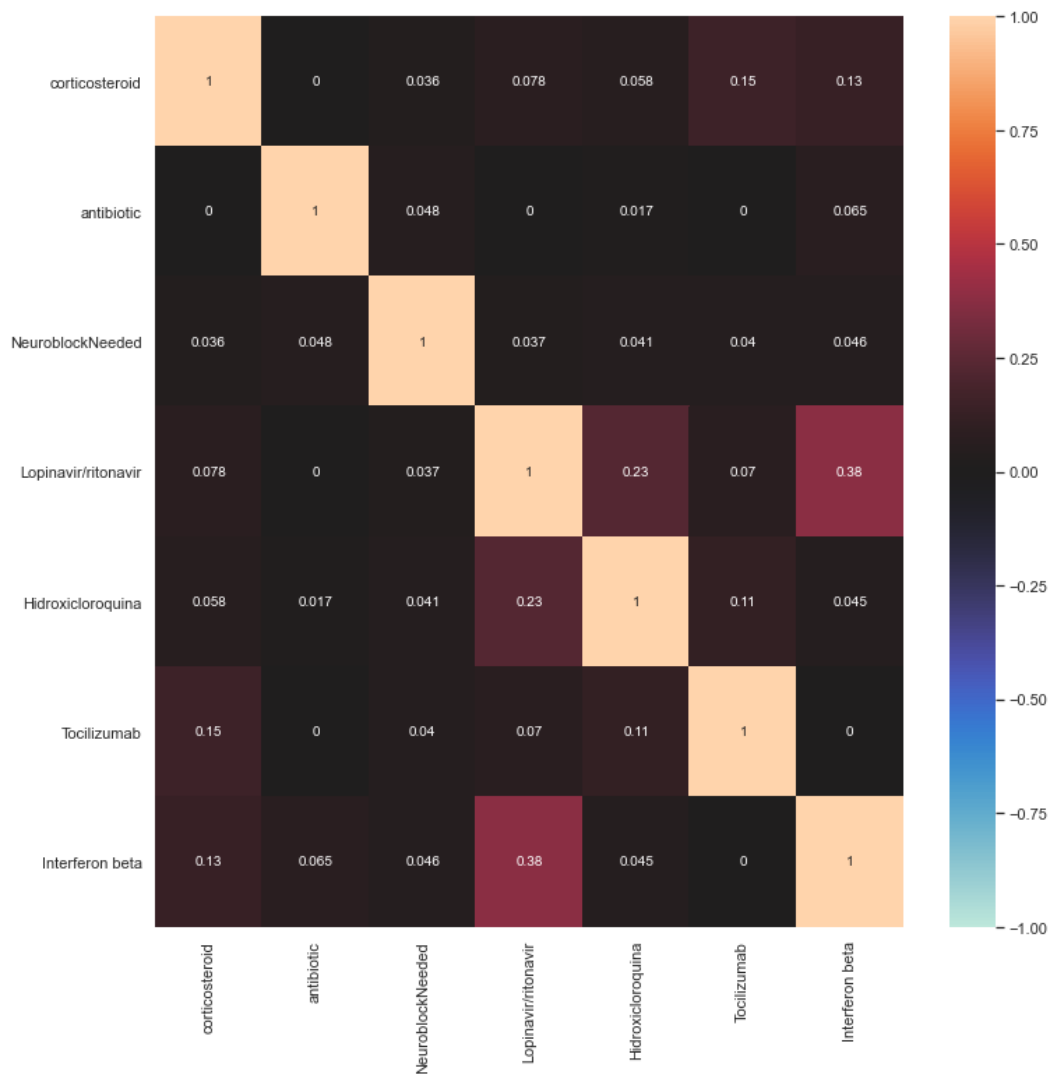


Figure 10.37: Correlation plot for treatment variables

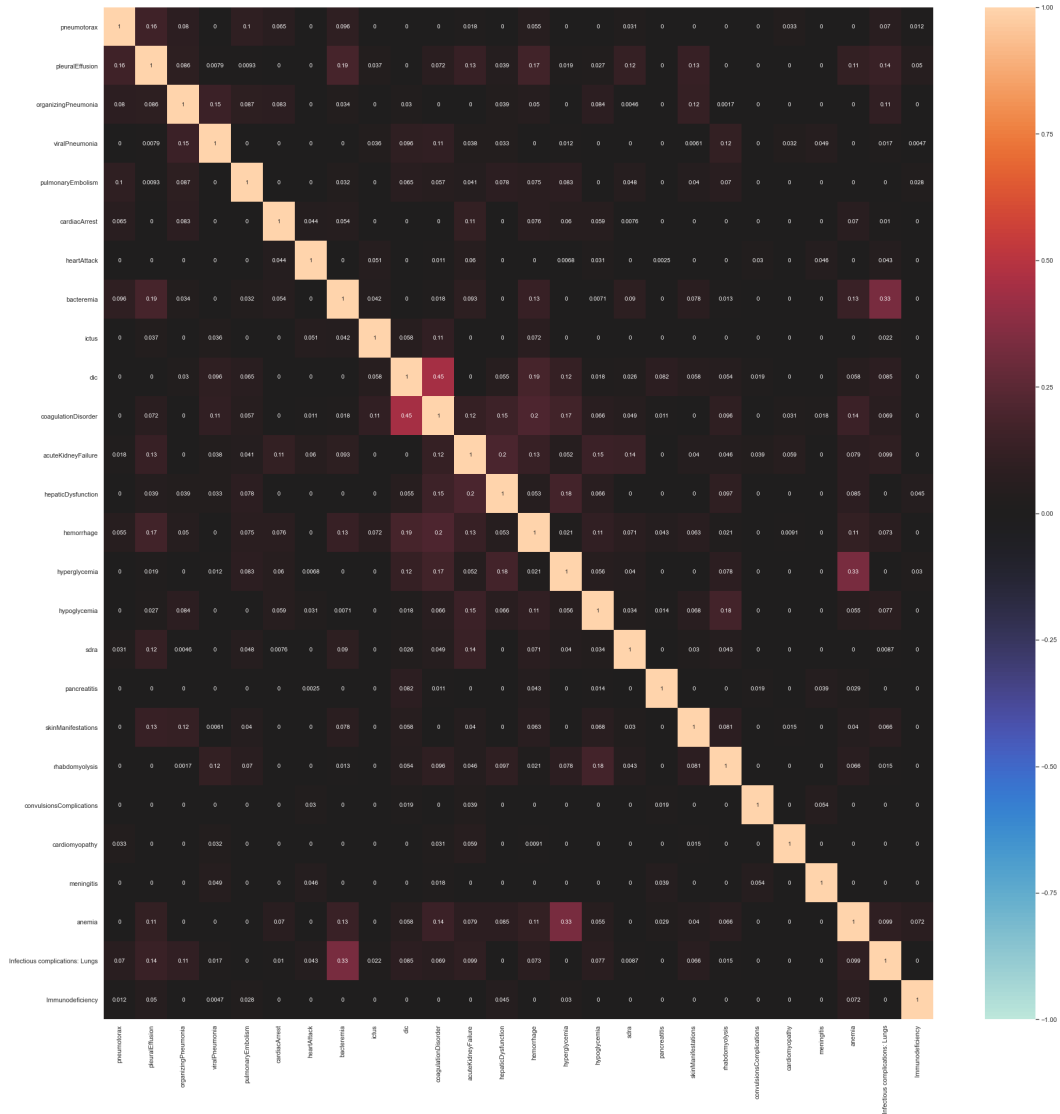


Figure 10.38: Correlation plot for complication variables

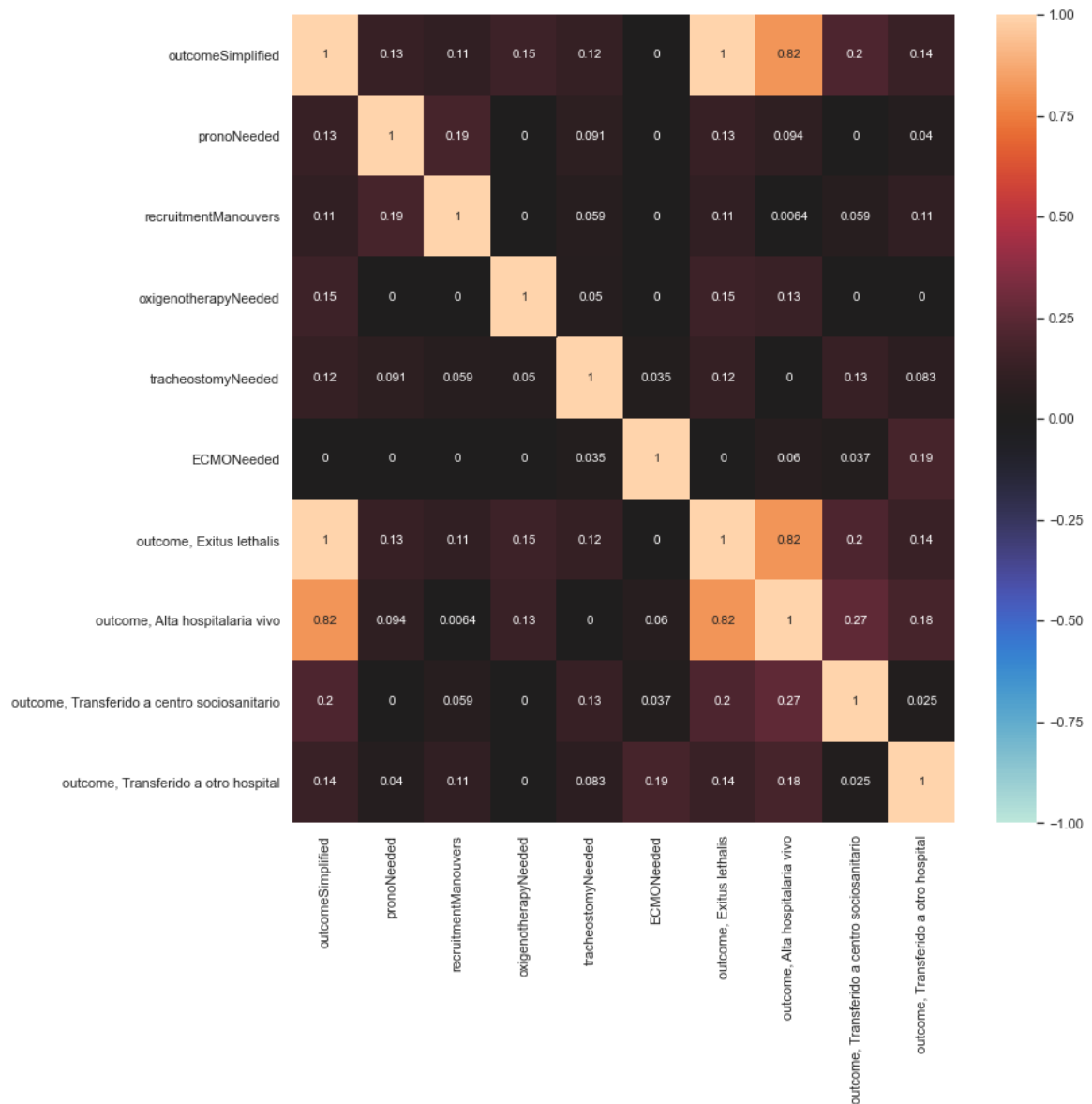


Figure 10.39: Correlation plot for outcome variables

	Variable	All (n=1140)	Alive (n=697)	Dead (n=443)	p value
0	Age	65.0 [56.0-71.0 IQR] (n=1140)	61.0 [53.0-69.0 IQR] (n=697)	69.0 [62.5-74.0 IQR] (n=443)	0.0 *
1	Aceinhibitors	201/1140 (17.63%)	108/697 (15.49%)	93/443 (20.99%)	0.021 *
2	Arb	189/1140 (16.58%)	113/697 (16.21%)	76/443 (17.16%)	0.684
3	Betablockers	106/1140 (9.3%)	51/697 (7.32%)	55/443 (12.42%)	0.005 *
4	Calciumchannelblockers	155/1140 (13.6%)	90/697 (12.91%)	65/443 (14.67%)	0.425
5	Diuretics	169/1140 (14.82%)	95/697 (13.63%)	74/443 (16.7%)	0.171
6	Statins	345/1134 (30.42%)	192/697 (27.55%)	153/437 (35.01%)	0.01 *
7	Streptococcusvaccine	90/635 (14.17%)	53/429 (12.35%)	37/206 (17.96%)	0.068
8	Fluvaccine	181/717 (25.24%)	107/478 (22.38%)	74/239 (30.96%)	0.014 *
9	Sex, male	816/1140 (71.58%)	481/697 (69.01%)	335/443 (75.62%)	0.018 *
10	Sex, female	324/1140 (28.42%)	216/697 (30.99%)	108/443 (24.38%)	0.018 *
11	Hospital admission-start symptoms	7.0 [5.0-10.0 IQR] (n=1120)	7.0 [5.0-10.0 IQR] (n=687)	7.0 [4.0-10.0 IQR] (n=433)	0.057

Figure 10.40: Characteristics of the population according to gender, age and previous medication at hospital admission

	Variable	All (n=1140)	Alive (n=697)	Dead (n=443)	p value
0	Hypertension	567/1139 (49.78%)	315/697 (45.19%)	252/442 (57.01%)	0.0 *
1	Obesity	383/1138 (33.66%)	254/696 (36.49%)	129/442 (29.19%)	0.012 *
2	Diabetes	260/1139 (22.83%)	155/697 (22.24%)	105/442 (23.76%)	0.563
3	Hematological	58/1139 (5.09%)	29/697 (4.16%)	29/442 (6.56%)	0.096
4	Renalchronic	60/1139 (5.27%)	26/697 (3.73%)	34/442 (7.69%)	0.004 *
5	Asthma	59/1139 (5.18%)	36/697 (5.16%)	23/442 (5.2%)	1.0
6	Hiv	6/1139 (0.53%)	4/697 (0.57%)	2/442 (0.45%)	1.0
7	Heartchronic	139/1139 (12.2%)	60/697 (8.61%)	79/442 (17.87%)	0.0 *
8	Pulmonarchronic	126/1139 (11.06%)	54/697 (7.75%)	72/442 (16.29%)	0.0 *
9	Dementia	5/1139 (0.44%)	2/697 (0.29%)	3/442 (0.68%)	0.382
10	Severehepatic	9/1139 (0.79%)	8/697 (1.15%)	1/442 (0.23%)	0.166
11	Smoker	63/757 (8.32%)	34/478 (7.11%)	29/279 (10.39%)	0.133

Figure 10.41: Characteristics of the population according to comorbidities

	Variable	All (n=1140)	Alive (n=697)	Dead (n=443)	p value
0	Fever	999/1125 (88.8%)	615/688 (89.39%)	384/437 (87.87%)	0.439
1	Drycough	755/1120 (67.41%)	459/684 (67.11%)	296/436 (67.89%)	0.794
2	Productivecough	148/1118 (13.24%)	93/684 (13.6%)	55/434 (12.67%)	0.717
3	Hemoptysiscough	18/1119 (1.61%)	12/685 (1.75%)	6/434 (1.38%)	0.808
4	Throatache	73/1109 (6.58%)	55/681 (8.08%)	18/428 (4.21%)	0.012 *
5	Nasalcongestion	45/1108 (4.06%)	29/679 (4.27%)	16/429 (3.73%)	0.755
6	Wheeze	121/1105 (10.95%)	73/681 (10.72%)	48/424 (11.32%)	0.767
7	Shortnessbreath	808/1131 (71.44%)	508/694 (73.2%)	300/437 (68.65%)	0.105
8	Chestpain	103/1118 (9.21%)	69/686 (10.06%)	34/432 (7.87%)	0.243
9	Musclepain	287/1109 (25.88%)	203/682 (29.77%)	84/427 (19.67%)	0.0 *
10	Articulationpain	196/1105 (17.74%)	130/678 (19.17%)	66/427 (15.46%)	0.125
11	Fatigue	377/1114 (33.84%)	239/684 (34.94%)	138/430 (32.09%)	0.363
12	Walkincapacity	20/1111 (1.8%)	10/681 (1.47%)	10/430 (2.33%)	0.355
13	Headache	89/1113 (8.0%)	64/683 (9.37%)	25/430 (5.81%)	0.041 *
14	Confusion	63/1123 (5.61%)	27/687 (3.93%)	36/436 (8.26%)	0.003 *
15	Syncope	33/1122 (2.94%)	20/686 (2.92%)	13/436 (2.98%)	1.0
16	Convulsionssymptoms	2/1123 (0.18%)	2/687 (0.29%)	0/436 (0.0%)	0.525
17	Abdominalpain	60/1122 (5.35%)	30/687 (4.37%)	30/435 (6.9%)	0.077
18	Nausea	106/1123 (9.44%)	67/687 (9.75%)	39/436 (8.94%)	0.677
19	Diarrhea	210/1124 (18.68%)	126/688 (18.31%)	84/436 (19.27%)	0.695
20	Anorexia	67/1113 (6.02%)	42/684 (6.14%)	25/429 (5.83%)	0.897
21	Smellloss	64/1047 (6.11%)	45/653 (6.89%)	19/394 (4.82%)	0.186
22	Tasteloss	63/1045 (6.03%)	48/651 (7.37%)	15/394 (3.81%)	0.022 *
23	Skinrashes	3/1120 (0.27%)	2/686 (0.29%)	1/434 (0.23%)	1.0
24	Skinulceras	1/1122 (0.09%)	1/686 (0.15%)	0/436 (0.0%)	1.0
25	Conjunctivitis	0/1121 (0.0%)	0/686 (0.0%)	0/435 (0.0%)	1.0
26	Lymphadenopathy	1/1068 (0.09%)	1/656 (0.15%)	0/412 (0.0%)	1.0
27	Bleeding	4/1124 (0.36%)	3/688 (0.44%)	1/436 (0.23%)	1.0

Figure 10.42: Characteristics of the population according to symptoms

	Variable	All (n=1140)	Alive (n=697)	Dead (n=443)	p value
0	Il6, pg/ml	114.0 [51.16-186.02 IQR] (n=219)	104.25 [45.89-188.67 IQR] (n=150)	125.86 [60.21-164.9 IQR] (n=69)	0.626
1	Vasopressorrequirement	751/1129 (66.52%)	453/693 (65.37%)	298/436 (68.35%)	0.331
2	Ecmorequirement	26/1130 (2.3%)	16/693 (2.31%)	10/437 (2.29%)	1.0
3	Breathingrate, rpm	24.0 [20.0-30.0 IQR] (n=982)	24.0 [20.0-30.0 IQR] (n=626)	24.0 [20.0-30.0 IQR] (n=356)	0.811
4	Heartrate, lpm	88.0 [72.0-103.0 IQR] (n=1033)	88.0 [70.0-101.0 IQR] (n=632)	89.0 [75.0-105.0 IQR] (n=401)	0.059
5	Temperature, °c	36.9 [36.0-37.8 IQR] (n=1021)	37.0 [36.0-37.8 IQR] (n=628)	36.9 [36.0-37.8 IQR] (n=393)	0.217
6	Hco3, mmol/l	24.3 [22.0-27.0 IQR] (n=1042)	24.7 [22.3-27.2 IQR] (n=638)	23.95 [21.18-26.9 IQR] (n=404)	0.003 *
7	Paco2, mmhg	43.4 [36.0-52.0 IQR] (n=1104)	42.95 [36.0-50.22 IQR] (n=678)	45.0 [37.0-55.0 IQR] (n=426)	0.001 *
8	Variablesldh, u/l	527.0 [413.0-701.0 IQR] (n=881)	505.0 [392.0-657.0 IQR] (n=533)	561.0 [442.0-739.25 IQR] (n=348)	0.0 *
9	Leucocytes, 10^9/l	8.9 [6.6-12.8 IQR] (n=1124)	8.7 [6.4-12.2 IQR] (n=689)	9.4 [6.91-13.6 IQR] (n=435)	0.005 *
10	Creatinine, mg/dl	0.86 [0.67-1.15 IQR] (n=1129)	0.81 [0.64-1.07 IQR] (n=693)	0.95 [0.75-1.29 IQR] (n=436)	0.0 *
11	Lymphocytes, 10^9/l	0.63 [0.42-0.9 IQR] (n=1120)	0.68 [0.47-0.95 IQR] (n=683)	0.6 [0.4-0.84 IQR] (n=437)	0.001 *
12	Crp, mg/dl	17.34 [9.2-26.4 IQR] (n=1001)	16.54 [9.0-25.62 IQR] (n=616)	18.52 [9.7-27.6 IQR] (n=385)	0.031 *
13	Procalcitonin, ng/ml	0.3 [0.15-0.8 IQR] (n=770)	0.26 [0.14-0.62 IQR] (n=461)	0.4 [0.2-1.02 IQR] (n=309)	0.0 *
14	Lactate, mg/dl	13.51 [9.91-18.02 IQR] (n=820)	12.61 [9.81-16.21 IQR] (n=496)	14.41 [10.81-20.72 IQR] (n=324)	0.0 *
15	Ddimer, mg/l	1.14 [0.6-3.67 IQR] (n=896)	1.0 [0.53-2.39 IQR] (n=559)	1.76 [0.7-6.08 IQR] (n=337)	0.0 *
16	Ferritin, ng/ml	1380.0 [795.0-2232.0 IQR] (n=397)	1321.5 [766.0-2104.25 IQR] (n=260)	1465.0 [864.0-2376.0 IQR] (n=137)	0.131
17	Urea, mg/dl	45.0 [32.0-64.0 IQR] (n=990)	41.0 [30.0-58.0 IQR] (n=603)	50.9 [37.0-68.0 IQR] (n=387)	0.0 *
18	Platelets, 10^9/l	227.5 [173.0-298.75 IQR] (n=1126)	236.0 [182.0-306.0 IQR] (n=688)	212.0 [163.25-283.5 IQR] (n=438)	0.0 *
19	Glucose, mg/dl	144.0 [115.0-194.55 IQR] (n=1083)	140.0 [110.0-182.0 IQR] (n=664)	153.0 [124.5-205.0 IQR] (n=419)	0.0 *
20	Variablesprothrombintime, seconds	13.4 [12.4-14.7 IQR] (n=799)	13.3 [12.4-14.5 IQR] (n=498)	13.6 [12.5-14.9 IQR] (n=301)	0.114
21	Totalbilirubin, mg/dl	0.63 [0.42-1.0 IQR] (n=1039)	0.66 [0.43-1.0 IQR] (n=636)	0.6 [0.4-1.0 IQR] (n=403)	0.628
22	Variablesast, u/l	52.0 [35.0-76.0 IQR] (n=852)	52.0 [34.0-75.0 IQR] (n=516)	52.4 [36.75-79.0 IQR] (n=336)	0.34
23	Variablesalt, u/l	39.0 [25.0-63.0 IQR] (n=1039)	40.0 [25.0-65.0 IQR] (n=643)	37.0 [23.0-58.75 IQR] (n=396)	0.044 *
24	Troponint, ng/ml	0.02 [0.01-0.04 IQR] (n=291)	0.01 [0.01-0.03 IQR] (n=161)	0.03 [0.01-0.09 IQR] (n=130)	0.0 *
25	Troponini, ng/ml	0.02 [0.01-0.06 IQR] (n=372)	0.01 [0.01-0.05 IQR] (n=234)	0.03 [0.01-0.1 IQR] (n=138)	0.001 *
26	Variablesntprobnp, pg/ml	482.5 [183.05-1292.5 IQR] (n=180)	324.5 [146.0-1120.75 IQR] (n=108)	905.3 [371.75-1663.0 IQR] (n=72)	0.0 *
27	Sofa	7.0 [5.0-8.0 IQR] (n=809)	7.0 [4.0-8.0 IQR] (n=500)	7.0 [5.0-9.0 IQR] (n=309)	0.0 *
28	Sofa_hemo	3.0 [0.0-4.0 IQR] (n=1067)	3.0 [0.0-4.0 IQR] (n=660)	3.0 [0.0-4.0 IQR] (n=407)	0.156
29	Apache	12.0 [9.0-15.0 IQR] (n=651)	11.0 [8.0-14.0 IQR] (n=417)	13.5 [11.0-17.0 IQR] (n=234)	0.0 *

Figure 10.43: Characteristics of the population at ICU admission

	Variable	All (n=1140)	Alive (n=697)	Dead (n=443)	p value
0	Il6, pg/ml	104.3 [43.0-196.93 IQR] (n=121)	81.03 [44.4-190.98 IQR] (n=92)	126.0 [41.31-197.3 IQR] (n=29)	0.464
1	Vasopressorrequirement	777/1130 (68.76%)	457/693 (65.95%)	320/437 (73.23%)	0.01 *
2	Ecmorequirement	32/1134 (2.82%)	18/696 (2.59%)	14/438 (3.2%)	0.583
3	Breathingrate, rpm	22.0 [19.0-25.0 IQR] (n=932)	22.0 [18.0-25.0 IQR] (n=591)	22.0 [20.0-26.0 IQR] (n=341)	0.049 *
4	Heartrate, lpm	80.0 [64.0-97.0 IQR] (n=939)	79.0 [62.0-94.0 IQR] (n=564)	80.0 [67.0-100.0 IQR] (n=375)	0.003 *
5	Temperature, °c	36.8 [36.0-37.5 IQR] (n=965)	36.85 [36.0-37.5 IQR] (n=596)	36.6 [36.0-37.5 IQR] (n=369)	0.162
6	Hco3, mmol/l	28.0 [24.9-31.3 IQR] (n=1046)	28.2 [25.35-31.4 IQR] (n=643)	27.3 [24.0-31.2 IQR] (n=403)	0.015 *
7	Paco2, mmhg	47.0 [41.0-54.0 IQR] (n=1113)	46.0 [40.0-51.0 IQR] (n=680)	50.0 [44.0-57.0 IQR] (n=433)	0.0 *
8	Variablesldh, u/l	418.0 [329.5-556.0 IQR] (n=795)	391.0 [314.4-508.0 IQR] (n=485)	455.0 [362.0-613.75 IQR] (n=310)	0.0 *
9	Leucocytes, 10^9/l	9.0 [6.66-12.04 IQR] (n=1119)	8.66 [6.3-11.59 IQR] (n=685)	9.48 [7.21-12.93 IQR] (n=434)	0.0 *
10	Creatinine, mg/dl	0.91 [0.66-1.39 IQR] (n=1123)	0.8 [0.61-1.16 IQR] (n=687)	1.11 [0.78-1.94 IQR] (n=436)	0.0 *
11	Lymphocytes, 10^9/l	0.66 [0.42-1.0 IQR] (n=1116)	0.7 [0.5-1.0 IQR] (n=682)	0.6 [0.38-0.9 IQR] (n=434)	0.0 *
12	Crp, mg/dl	9.93 [3.2-22.3 IQR] (n=945)	9.12 [2.9-20.44 IQR] (n=582)	10.7 [3.79-23.34 IQR] (n=363)	0.072
13	Procalcitonin, ng/ml	0.37 [0.15-1.22 IQR] (n=533)	0.28 [0.11-0.77 IQR] (n=322)	0.65 [0.24-1.87 IQR] (n=211)	0.0 *
14	Lactate, mg/dl	15.31 [11.71-20.0 IQR] (n=825)	15.31 [10.81-19.75 IQR] (n=512)	16.21 [12.61-20.72 IQR] (n=313)	0.0 *
15	Ddimer, mg/l	2.32 [1.03-6.25 IQR] (n=784)	1.99 [0.96-4.69 IQR] (n=478)	3.61 [1.38-8.87 IQR] (n=306)	0.0 *
16	Ferritin, ng/ml	1339.0 [837.5-2176.5 IQR] (n=358)	1319.5 [830.75-2116.25 IQR] (n=248)	1350.0 [839.0-2239.05 IQR] (n=110)	0.695
17	Urea, mg/dl	57.53 [39.0-85.0 IQR] (n=1002)	52.0 [35.8-76.0 IQR] (n=613)	69.0 [45.8-103.2 IQR] (n=389)	0.0 *
18	Platelets, 10^9/l	254.0 [188.0-324.25 IQR] (n=1116)	274.0 [206.0-339.5 IQR] (n=683)	219.0 [167.0-293.0 IQR] (n=433)	0.0 *
19	Glucose, mg/dl	154.0 [125.0-198.75 IQR] (n=1086)	149.0 [120.0-191.0 IQR] (n=667)	164.0 [134.0-203.74 IQR] (n=419)	0.0 *
20	Variablesprothrombintime, seconds	13.1 [12.1-14.6 IQR] (n=771)	13.15 [12.1-14.5 IQR] (n=480)	13.0 [12.0-15.0 IQR] (n=291)	0.776
21	Totalbilirubin, mg/dl	0.76 [0.43-1.3 IQR] (n=975)	0.77 [0.44-1.3 IQR] (n=601)	0.76 [0.42-1.37 IQR] (n=374)	0.793
22	Variablesast, u/l	43.2 [27.0-71.0 IQR] (n=877)	45.0 [26.0-74.5 IQR] (n=535)	43.0 [29.0-68.0 IQR] (n=342)	0.799
23	Variablesalt, u/l	40.0 [24.0-67.0 IQR] (n=1010)	44.0 [25.0-77.0 IQR] (n=624)	36.0 [23.0-58.0 IQR] (n=386)	0.0 *
24	Troponint, ng/ml	0.02 [0.01-0.06 IQR] (n=211)	0.02 [0.01-0.04 IQR] (n=114)	0.02 [0.01-0.1 IQR] (n=97)	0.0 *
25	Troponini, ng/ml	0.02 [0.01-0.05 IQR] (n=225)	0.01 [0.0-0.03 IQR] (n=146)	0.03 [0.01-0.31 IQR] (n=79)	0.0 *
26	Variablesntprobnp, pg/ml	488.2 [179.1-1610.5 IQR] (n=115)	348.4 [143.0-1087.0 IQR] (n=71)	646.0 [229.72-2042.25 IQR] (n=44)	0.024 *
27	Sofa	7.0 [5.0-8.0 IQR] (n=691)	6.0 [4.0-8.0 IQR] (n=437)	7.0 [5.25-9.75 IQR] (n=254)	0.0 *
28	Sofa_hemo	3.0 [0.75-4.0 IQR] (n=1028)	3.0 [0.0-4.0 IQR] (n=636)	3.0 [1.0-4.0 IQR] (n=392)	0.0 *
29	Apache	10.0 [8.0-13.0 IQR] (n=579)	9.0 [7.0-12.0 IQR] (n=365)	12.0 [9.0-17.0 IQR] (n=214)	0.0 *

Figure 10.44: Characteristics of the population at 3rd of ICU admission

	Variable	All (n=1140)	Alive (n=697)	Dead (n=443)	p value
0	Compliance	37.02 [29.01-50.0 IQR] (n=416)	37.14 [30.0-50.0 IQR] (n=257)	35.77 [27.89-50.0 IQR] (n=159)	0.291
1	Rass	-5.0 [-5.0--4.0 IQR] (n=893)	-5.0 [-5.0--4.0 IQR] (n=545)	-5.0 [-5.0--4.0 IQR] (n=348)	-
2	Sas	2.0 [1.0-2.0 IQR] (n=116)	2.0 [1.0-2.0 IQR] (n=65)	2.0 [1.0-2.0 IQR] (n=51)	0.95
3	Fr, rpm	20.0 [18.0-24.0 IQR] (n=1026)	20.0 [18.0-24.0 IQR] (n=627)	20.0 [18.0-24.0 IQR] (n=399)	0.529
4	Drivingpressure	12.0 [9.58-15.0 IQR] (n=432)	12.0 [9.0-14.0 IQR] (n=268)	12.0 [9.9-16.0 IQR] (n=164)	0.207
5	Plateaupressure, cmh ²	25.0 [22.0-28.0 IQR] (n=442)	25.0 [21.0-28.0 IQR] (n=274)	25.0 [22.0-28.0 IQR] (n=168)	0.089
6	Peep, cmh ²	12.0 [10.0-14.0 IQR] (n=1059)	12.0 [10.0-14.0 IQR] (n=646)	12.0 [10.0-14.0 IQR] (n=413)	0.822
7	Ventilatoryratio	1.69 [1.33-2.19 IQR] (n=802)	1.65 [1.3-2.07 IQR] (n=497)	1.83 [1.4-2.33 IQR] (n=305)	0.001 *
8	Paco2, mmhg	43.0 [36.0-52.0 IQR] (n=1077)	42.0 [36.0-50.0 IQR] (n=660)	45.0 [36.6-55.0 IQR] (n=417)	0.001 *
9	Imvrequirement	1113/1116 (99.73%)	680/682 (99.71%)	433/434 (99.77%)	1.0
10	Nimvrequirement	101/1108 (9.12%)	57/679 (8.39%)	44/429 (10.26%)	0.335
11	Pafi, mmhg	115.9 [79.9-169.32 IQR] (n=1052)	118.0 [82.75-174.0 IQR] (n=651)	112.5 [74.0-157.4 IQR] (n=401)	0.033 *
12	Variablesioxigensaturation, %	94.95 [90.0-97.4 IQR] (n=1040)	95.0 [90.4-97.5 IQR] (n=633)	94.0 [89.0-97.0 IQR] (n=407)	0.007 *
13	Sofa	7.0 [5.0-8.0 IQR] (n=784)	7.0 [4.0-8.0 IQR] (n=487)	7.0 [5.0-9.0 IQR] (n=297)	0.0 *
14	Sofa_hemo	3.0 [0.0-4.0 IQR] (n=1042)	3.0 [0.0-4.0 IQR] (n=642)	3.0 [0.0-4.0 IQR] (n=400)	0.101
15	Apache	12.0 [9.0-15.0 IQR] (n=634)	11.0 [8.0-14.0 IQR] (n=404)	14.0 [11.0-17.0 IQR] (n=230)	0.0 *
16	Ventilatoryratiomodified	1.72 [1.34-2.24 IQR] (n=916)	1.67 [1.31-2.1 IQR] (n=568)	1.86 [1.41-2.36 IQR] (n=348)	0.001 *

Figure 10.45: Characteristics of the population at the beginning of mechanical ventilation phase

	Variable	All (n=1140)	Alive (n=697)	Dead (n=443)	p value
0	Compliance	30.5 [20.43-44.15 IQR] (n=159)	42.33 [33.26-111.4 IQR] (n=51)	26.64 [17.61-36.69 IQR] (n=108)	0.0 *
1	Rass	-2.0 [-5.0-0.0 IQR] (n=766)	0.0 [-1.0-0.0 IQR] (n=435)	-5.0 [-5.0--4.0 IQR] (n=331)	0.673
2	Sas	2.0 [2.0-4.0 IQR] (n=81)	4.0 [4.0-4.0 IQR] (n=34)	2.0 [1.0-2.0 IQR] (n=47)	0.0 *
3	Fr, rpm	22.0 [18.0-26.0 IQR] (n=624)	20.0 [17.0-24.0 IQR] (n=278)	24.0 [20.0-27.0 IQR] (n=346)	0.0 *
4	Drivingpressure	14.0 [9.0-20.0 IQR] (n=160)	10.0 [5.0-13.22 IQR] (n=52)	16.45 [12.0-22.0 IQR] (n=108)	0.0 *
5	Plateaupressure, cmh ²	24.95 [20.0-30.0 IQR] (n=160)	19.5 [14.0-22.0 IQR] (n=50)	28.0 [24.0-32.08 IQR] (n=110)	0.0 *
6	Peep, cmh ²	8.0 [6.0-10.0 IQR] (n=897)	7.0 [6.0-8.0 IQR] (n=508)	10.0 [8.0-12.0 IQR] (n=389)	0.0 *
7	Ventilatoryratio	2.26 [1.66-3.07 IQR] (n=381)	1.67 [1.38-2.15 IQR] (n=145)	2.72 [2.11-3.5 IQR] (n=236)	0.0 *
8	Paco2, mmhg	45.0 [39.0-56.0 IQR] (n=955)	42.0 [37.0-46.0 IQR] (n=580)	59.0 [47.35-69.0 IQR] (n=375)	0.0 *
9	Imvrequirement	1017/1126 (90.32%)	584/690 (84.64%)	433/436 (99.31%)	0.0 *
10	Nimvrequirement	88/1124 (7.83%)	82/688 (11.92%)	6/436 (1.38%)	0.0 *
11	Pafi, mmhg	186.0 [107.5-248.0 IQR] (n=887)	228.57 [185.67-277.05 IQR] (n=515)	101.55 [73.94-153.39 IQR] (n=372)	0.0 *
12	Variablesioxigensaturation, %	96.0 [93.0-98.0 IQR] (n=1009)	97.0 [95.0-98.0 IQR] (n=618)	93.3 [89.0-96.0 IQR] (n=391)	0.0 *
13	Sofa	4.0 [2.0-8.0 IQR] (n=511)	3.0 [2.0-4.0 IQR] (n=305)	8.5 [6.0-11.0 IQR] (n=206)	0.0 *
14	Sofa_hemo	0.0 [0.0-4.0 IQR] (n=955)	0.0 [0.0-0.0 IQR] (n=574)	4.0 [1.0-4.0 IQR] (n=381)	0.0 *
15	Apache	12.0 [8.0-17.0 IQR] (n=458)	9.0 [7.0-12.0 IQR] (n=300)	18.0 [14.0-22.0 IQR] (n=158)	0.0 *
16	Ventilatoryratiomodified	2.12 [1.61-2.89 IQR] (n=584)	1.75 [1.39-2.25 IQR] (n=301)	2.6 [2.03-3.4 IQR] (n=283)	0.0 *

Figure 10.46: Characteristics of the population at the end of mechanical ventilation phase

	Variable	All (n=1140)	Alive (n=697)	Dead (n=443)	p value
0	Pneumotorax	120/1140 (10.53%)	58/697 (8.32%)	62/443 (14.0%)	0.003 *
1	Pleuraleffusion	140/1139 (12.29%)	87/696 (12.5%)	53/443 (11.96%)	0.853
2	Organizingpneumonia	57/1114 (5.12%)	41/685 (5.99%)	16/429 (3.73%)	0.124
3	Viralpneumonia	71/1140 (6.23%)	39/697 (5.6%)	32/443 (7.22%)	0.314
4	Pulmonaryembolism	100/1105 (9.05%)	74/684 (10.82%)	26/421 (6.18%)	0.009 *
5	Cardiacarrest	92/1139 (8.08%)	13/696 (1.87%)	79/443 (17.83%)	0.0 *
6	Heartattack	11/1140 (0.96%)	5/697 (0.72%)	6/443 (1.35%)	0.355
7	Bacteremia	458/1137 (40.28%)	267/696 (38.36%)	191/441 (43.31%)	0.107
8	Ictus	36/1137 (3.17%)	23/697 (3.3%)	13/440 (2.95%)	0.862
9	Dic	76/1124 (6.76%)	39/688 (5.67%)	37/436 (8.49%)	0.069
10	Coagulationdisorder	305/1138 (26.8%)	167/696 (23.99%)	138/442 (31.22%)	0.009 *
11	Chronickidneydisease	499/1140 (43.77%)	237/697 (34.0%)	262/443 (59.14%)	0.0 *
12	Hepaticdysfunction	380/1137 (33.42%)	222/697 (31.85%)	158/440 (35.91%)	0.175
13	Hemorrhage	121/1138 (10.63%)	54/695 (7.77%)	67/443 (15.12%)	0.0 *
14	Hyperglycemia	828/1140 (72.63%)	508/697 (72.88%)	320/443 (72.23%)	0.838
15	Hypoglycemia	52/1140 (4.56%)	28/697 (4.02%)	24/443 (5.42%)	0.308
16	Sdra	980/1138 (86.12%)	588/696 (84.48%)	392/442 (88.69%)	0.053
17	Pancreatitis	14/1140 (1.23%)	8/697 (1.15%)	6/443 (1.35%)	0.787
18	Skinmanifestations	85/1137 (7.48%)	68/696 (9.77%)	17/441 (3.85%)	0.0 *
19	Rhabdomyolysis	47/1133 (4.15%)	27/694 (3.89%)	20/439 (4.56%)	0.647
20	Convulsionscomplications	9/1140 (0.79%)	6/697 (0.86%)	3/443 (0.68%)	1.0
21	Cardiomyopathy	27/1139 (2.37%)	19/697 (2.73%)	8/442 (1.81%)	0.425
22	Meningitis	6/1138 (0.53%)	5/697 (0.72%)	1/441 (0.23%)	0.414
23	Anemia	830/1140 (72.81%)	514/697 (73.74%)	316/443 (71.33%)	0.376
24	Corticosteroid	851/1113 (76.46%)	520/688 (75.58%)	331/425 (77.88%)	0.424
25	Antibiotic	1123/1133 (99.12%)	686/695 (98.71%)	437/438 (99.77%)	0.099
26	Antiviral	1096/1131 (96.91%)	675/694 (97.26%)	421/437 (96.34%)	0.384
27	Pulmones	484/1140 (42.46%)	298/697 (42.75%)	186/443 (41.99%)	0.806
28	Pulmones_hemo	23/1140 (2.02%)	9/697 (1.29%)	14/443 (3.16%)	0.032 *

Figure 10.47: Characteristics of the population according to treatments and complications during hospital stay

	Variable	All (n=1140)	Alive (n=697)	Dead (n=443)	p value
0	Timeinhospital	30.0 [18.0-48.0 IQR] (n=1140)	38.0 [25.0-57.0 IQR] (n=697)	19.0 [11.0-30.0 IQR] (n=443)	0.0 *
1	Timeinicu	20.0 [11.0-32.0 IQR] (n=1138)	22.0 [13.0-36.0 IQR] (n=697)	16.0 [9.0-26.0 IQR] (n=441)	0.0 *
2	Timeinimv	15.0 [9.0-27.0 IQR] (n=1131)	15.0 [10.0-28.0 IQR] (n=692)	15.0 [8.0-26.0 IQR] (n=439)	0.104
3	Prononeeded	883/1129 (78.21%)	512/693 (73.88%)	371/436 (85.09%)	0.0 *
4	Recruitmentmanouvers	662/1083 (61.13%)	376/662 (56.8%)	286/421 (67.93%)	0.0 *
5	Neuroblockneeded	973/1133 (85.88%)	571/695 (82.16%)	402/438 (91.78%)	0.0 *
6	Alive28days	765/1129 (67.76%)	688/688 (100.0%)	77/441 (17.46%)	0.0 *
7	Imvdays	16.0 [10.0-27.0 IQR] (n=1115)	16.0 [10.0-27.0 IQR] (n=680)	15.0 [9.0-25.0 IQR] (n=435)	0.103
8	Nimvdays	2.0 [1.0-5.0 IQR] (n=269)	3.0 [1.0-6.0 IQR] (n=196)	2.0 [1.0-4.0 IQR] (n=73)	0.183
9	Outcome, exitus lethalis	443/1140 (38.86%)	0/697 (0.0%)	443/443 (100.0%)	0.0 *
10	Outcome, alta hospitalaria vivo	585/1140 (51.32%)	585/697 (83.93%)	0/443 (0.0%)	0.0 *
11	Outcome, transferido a centro sociosanitario	74/1140 (6.49%)	74/697 (10.62%)	0/443 (0.0%)	0.0 *
12	Outcome, transferido a otro hospital	38/1140 (3.33%)	38/697 (5.45%)	0/443 (0.0%)	0.0 *
13	Reintubationneeded, no	137/149 (91.95%)	76/85 (89.41%)	61/64 (95.31%)	0.235
14	Reintubationneeded, sí, por fallo en la extubación	10/149 (6.71%)	7/85 (8.24%)	3/64 (4.69%)	0.516
15	Reintubationneeded, sí, por autoextubación	2/149 (1.34%)	2/85 (2.35%)	0/64 (0.0%)	0.507

Figure 10.48: Characteristics of the population according to complementary therapies and outcome

variable	# missing values	imputed	normal range
apache_day3	399		[0,inf]
sofa_day3	336		[0,inf]
apache_ICU	334		[0,inf]
variablesProthrombinTime_day3	314		[10.1,53.0]
procalcitonin_ICU	302		[0.01,150.0]
variablesProthrombinTime_ICU	297		[10.1,53.0]
dDimer_day3	284		[0,inf]
variablesLdh_day3	283		[77.0,3582.0]
lactate_ICU	257		[3.3,105.0]
lactate_day3	250		[3.3,105.0]
variablesAst_ICU	247		[6.0,5275.0]
ventilatoryRatio_ICU	235		[0.0,10.0]
ventilatoryRatio_day3	235		[0.0,10.0]
sofa_ICU	233		[0,inf]
variablesLdh_ICU	213		[77.0,3582.0]
variablesAst_day3	212		[6.0,5275.0]
dDimer_ICU	190		[0,inf]
rass_ICU	156		[-5.0,4.0]
breathingRate_day3	147		[6.0,45.0]
crp_day3	147		[0.2,70.0]

Figure 10.49: Number of missing values to be imputed per variable [1/3]

variable	# missing values imputed	normal range
heartrate_day3	144	[44.0,227.0]
ventilatoryRatioModified_ICU	129	[0.0,10.0]
totalBilirubin_day3	119	[0.1,27.3]
urea_ICU	117	[0,inf]
rass_day3	115	[-5.0,4.0]
breathingRate_ICU	112	[6.0,45.0]
ventilatoryRatioModified_day3	109	[0.0,10.0]
urea_day3	107	[0,inf]
crp_ICU	107	[0.2,70.0]
temperature_day3	106	[34.5,42.0]
variablesAlt_day3	80	[2.0,9423.0]
totalBilirubin_ICU	77	[0.1,27.3]
heartrate_ICU	76	[44.0,227.0]
temperature_ICU	75	[34.5,42.0]
variablesAlt_ICU	71	[2.0,9423.0]
fr_day3	64	[10.0,50.0]
sofa_hemo_day3	63	[0,inf]
HCO3_ICU	61	[14.0,40.1]
HCO3_day3	56	[14.0,40.1]
variablesOxygenSaturation_day3	55	[30.0,100.0]

Figure 10.50: Number of missing values to be imputed per variable [2/3]

variable	# missing values imputed	normal range
variablesOxygenSaturation_ICU	49	[30.0,100.0]
fr_ICU	49	[10.0,50.0]
glucose_ICU	36	[13.0,679.0]
sofa_hemo_ICU	34	[0,inf]
PaFi_ICU	33	[40.0,500.0]
glucose_day3	31	[13.0,679.0]
peep_ICU	29	[0.0,22.0]
peep_day3	22	[0.0,22.0]
paCO2_ICU	17	[10.0,100.0]
PaFi_day3	16	[40.0,500.0]
paCO2_day3	10	[10.0,100.0]
Hospital Admission-Start Symptoms	9	[0,inf]
lymphocytes_ICU	8	[0.0,30.0]
platelets_day3	8	[13.0,1200.0]
lymphocytes_day3	7	[0.0,30.0]
leucocytes_ICU	7	[0.27,80.32]
leucocytes_day3	3	[0.27,80.32]
platelets_ICU	3	[13.0,1200.0]
creatinine_day3	2	[0.15,12.0]
creatinine_ICU	1	[0.15,12.0]
timeInICU	1	[0,inf]

Figure 10.51: Number of missing values to be imputed per variable [3/3]

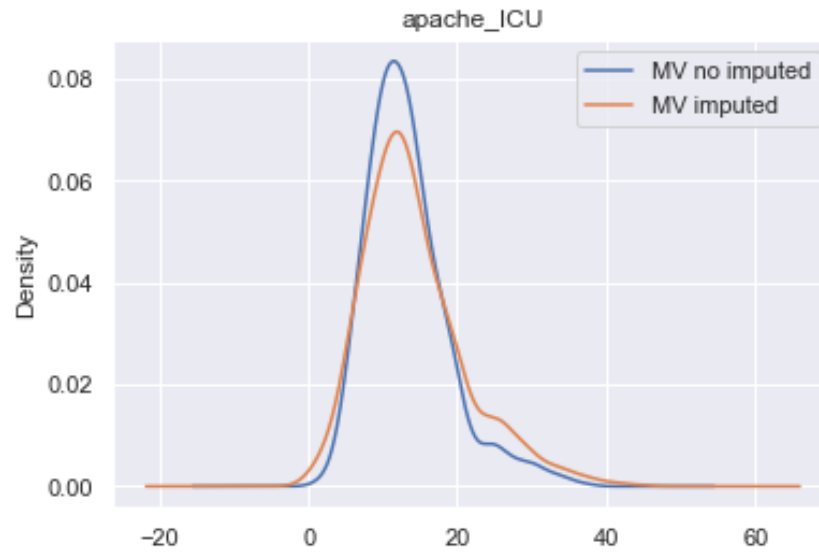


Figure 10.52: Apache score (ICU) density plot

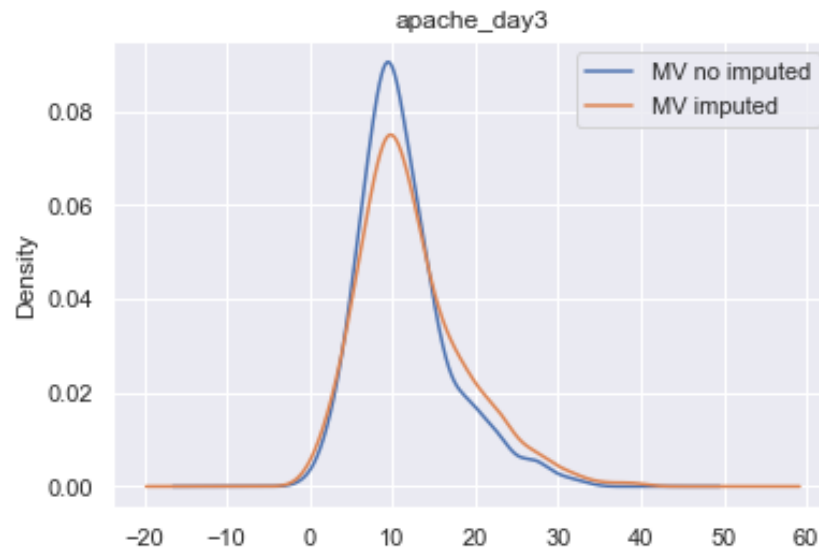


Figure 10.53: Apache score (3rd day ICU) density plot

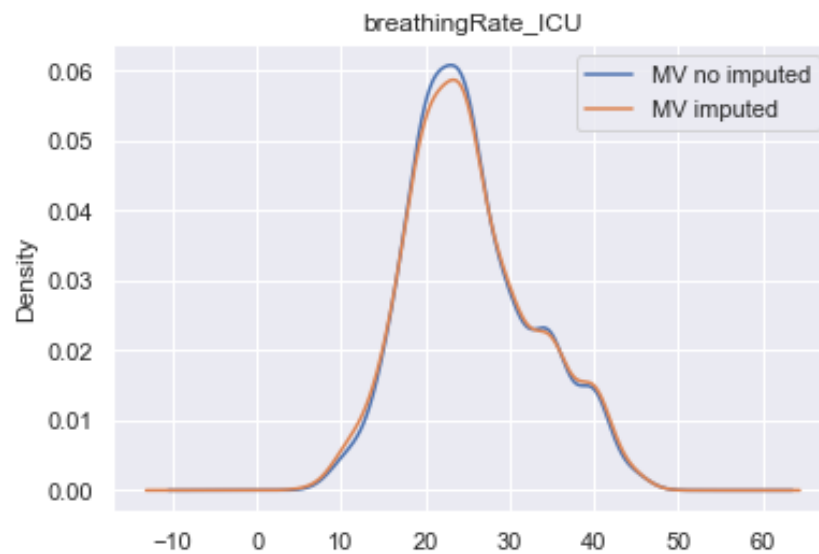


Figure 10.54: Breathing rate (ICU) density plot

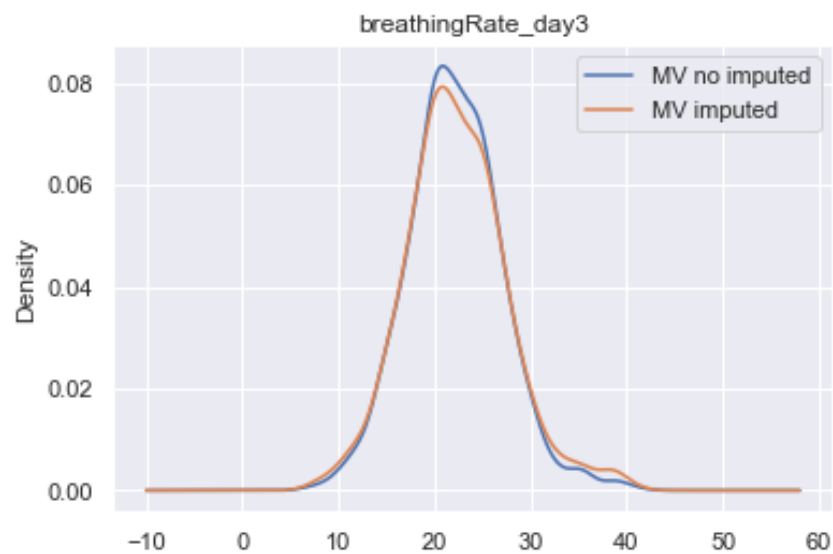


Figure 10.55: Breathing rate (3rd day ICU) density plot

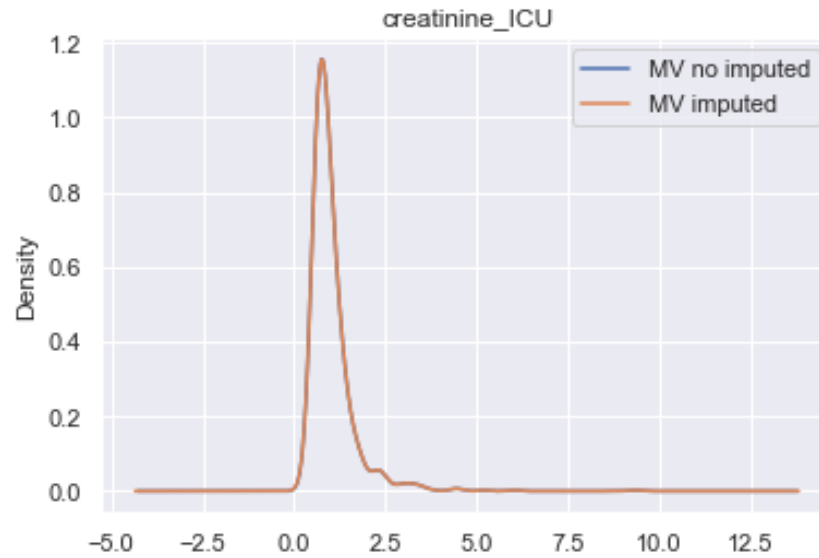


Figure 10.56: Creatinine (ICU) density plot

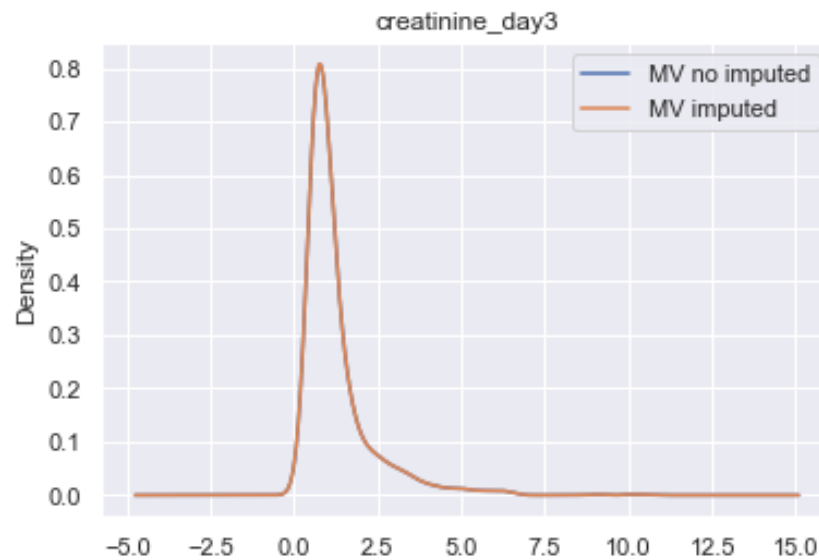


Figure 10.57: Creatinine (3rd day ICU) density plot

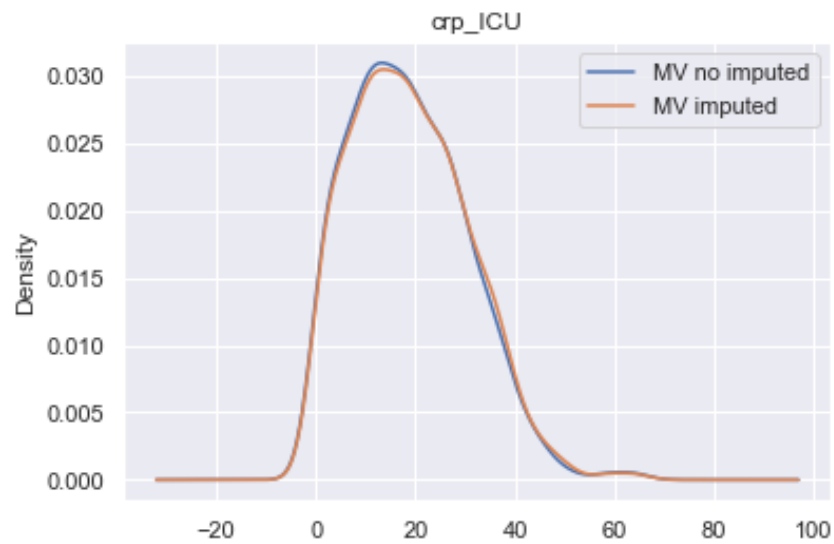


Figure 10.58: CRP (ICU) density plot

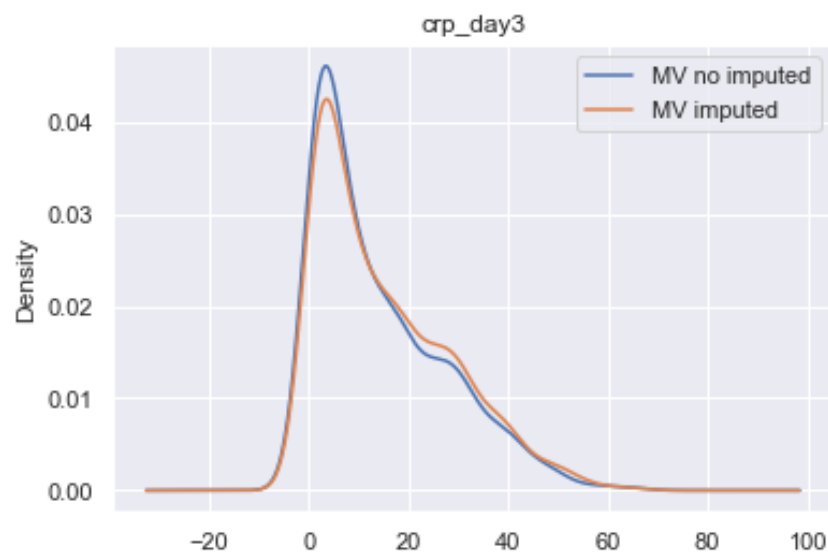


Figure 10.59: CRP (3rd day ICU) density plot

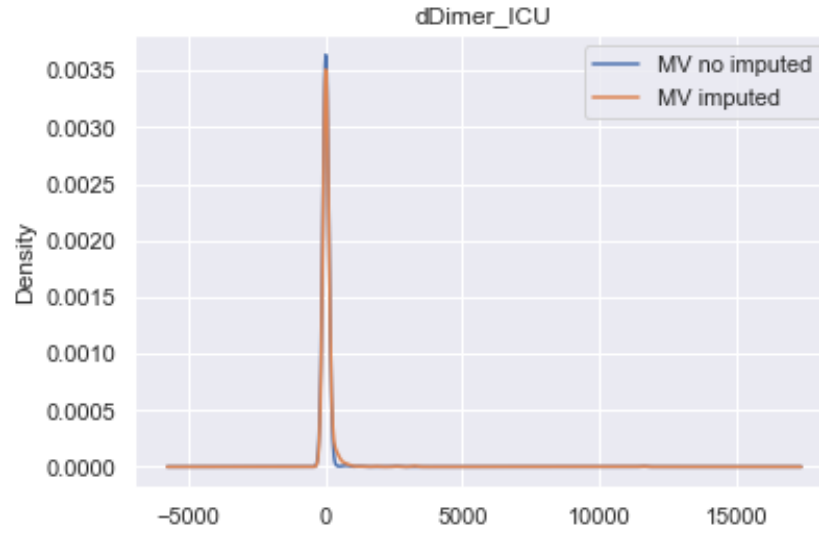


Figure 10.60: D-dimer (ICU) density plot

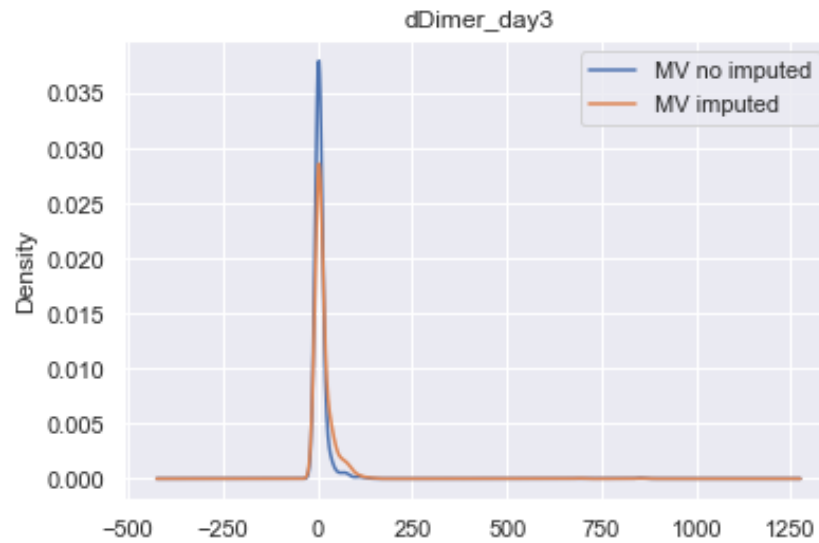


Figure 10.61: D-dimer (3rd day ICU) density plot

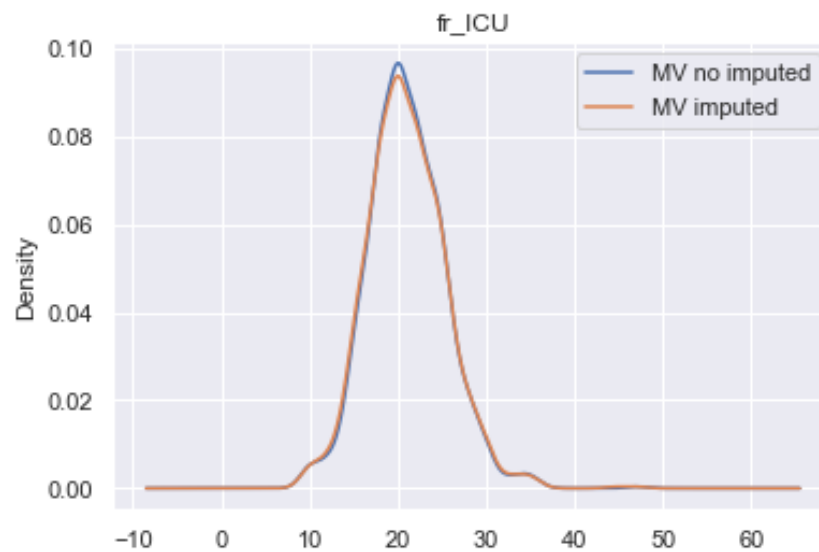


Figure 10.62: Regulated respiratory rate (FR) (ICU) density plot

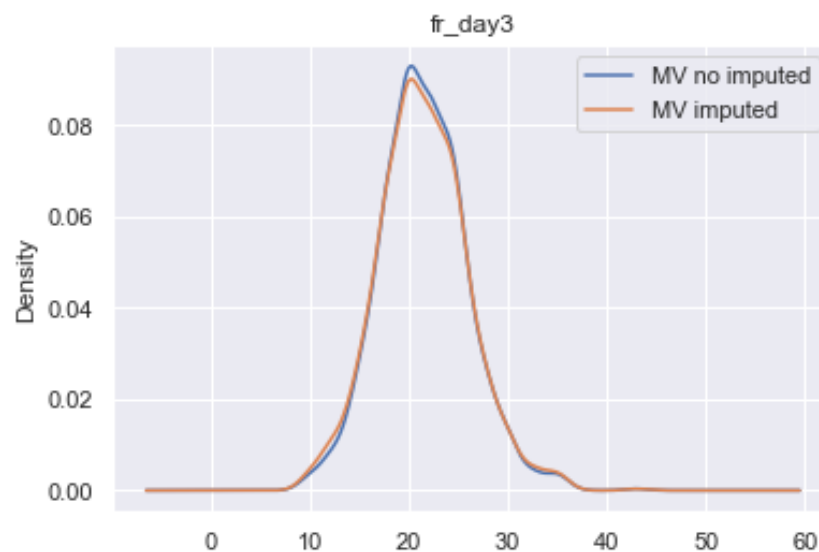


Figure 10.63: Regulated respiratory rate (FR) (3rd day ICU) density plot

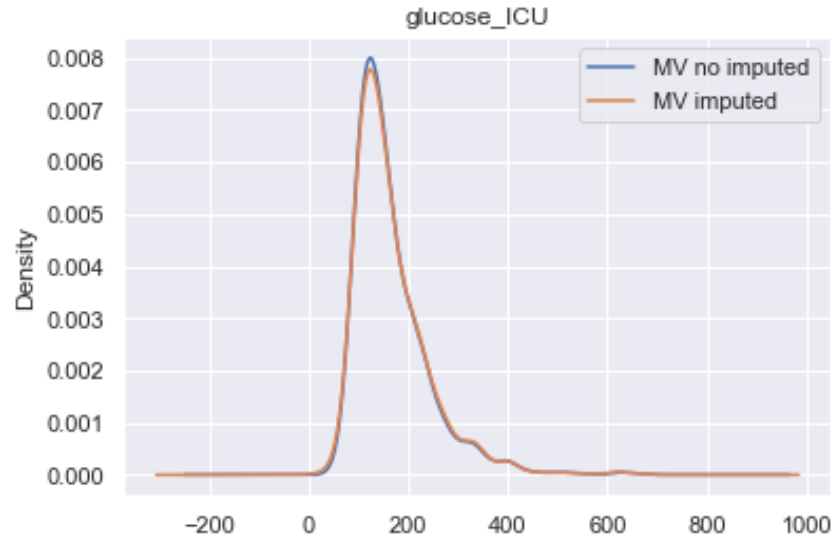


Figure 10.64: Glucose (ICU) density plot

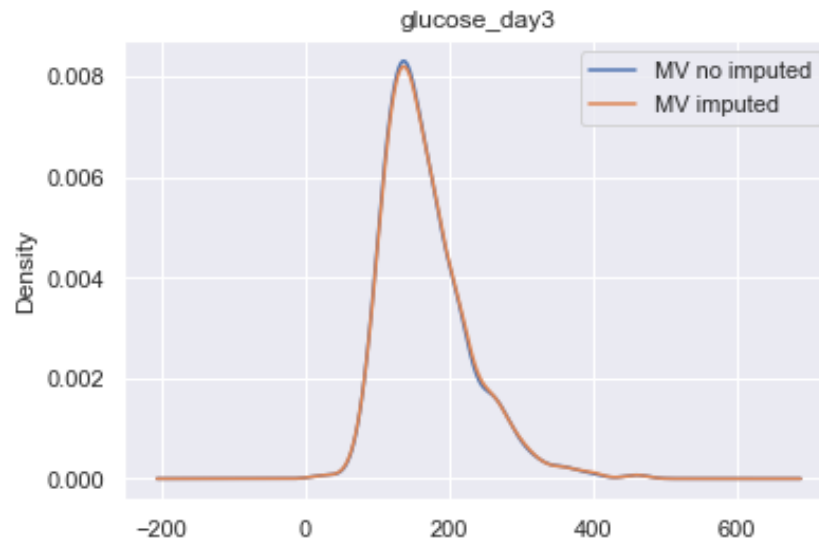


Figure 10.65: Glucose (3rd day ICU) density plot

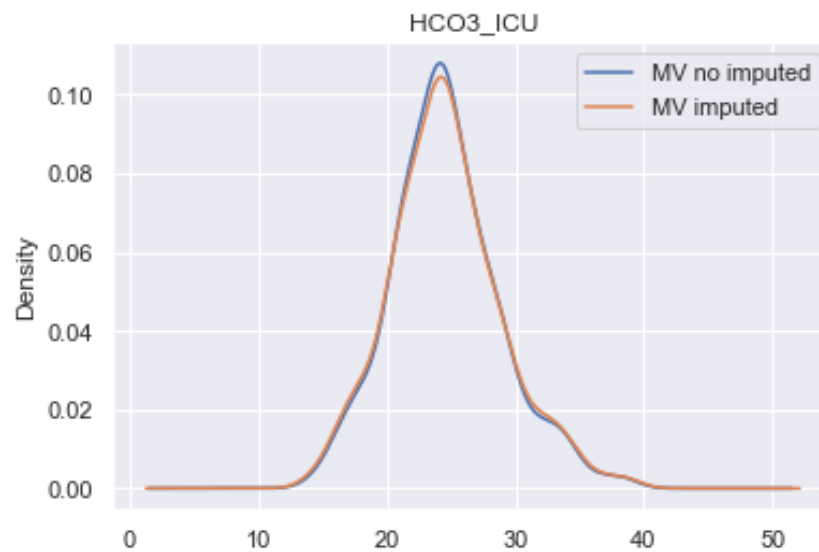


Figure 10.66: HCO3 (ICU) density plot

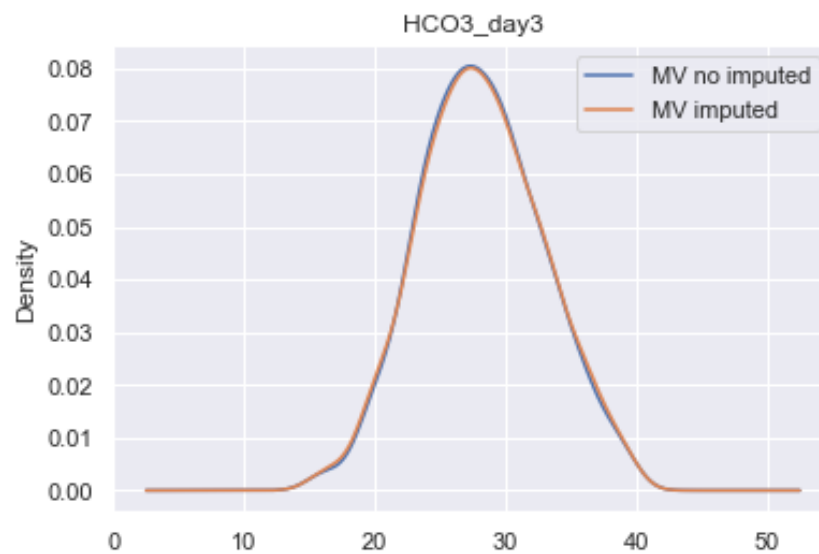


Figure 10.67: HCO3 (3rd day ICU) density plot

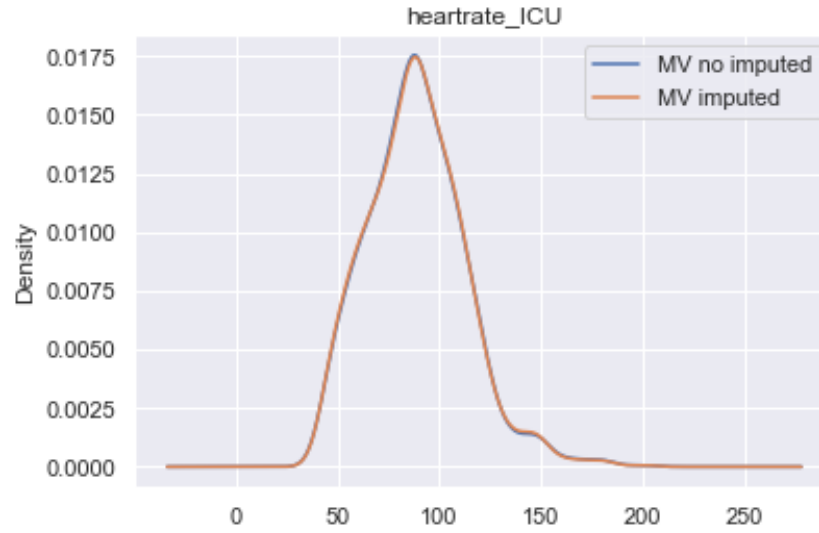


Figure 10.68: Heart rate (ICU) density plot

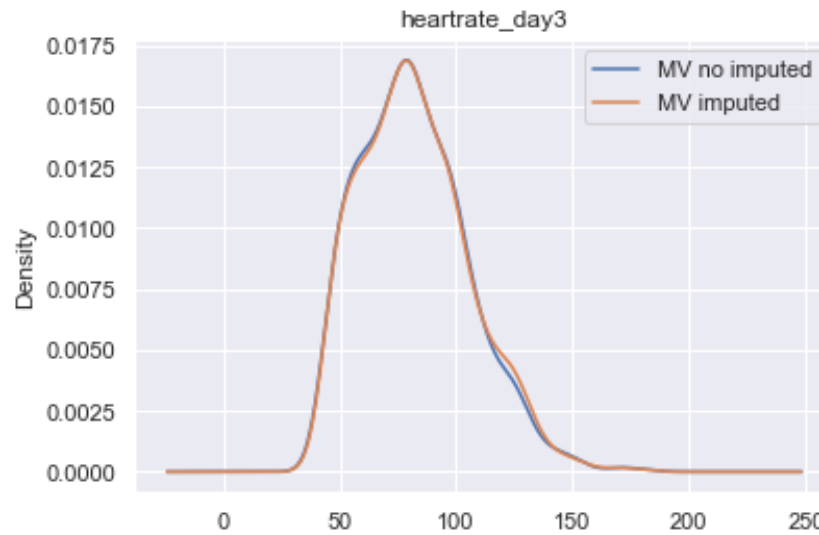


Figure 10.69: Heart rate (3rd day ICU) density plot

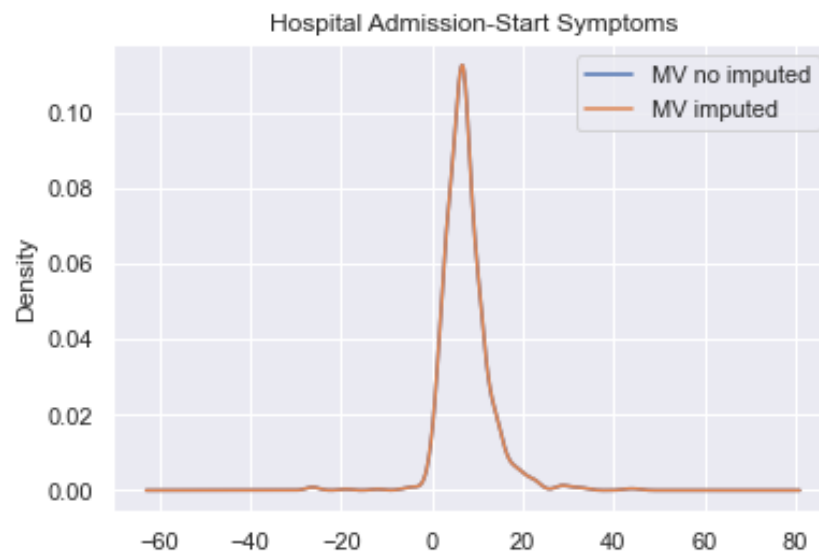


Figure 10.70: Time between symptoms appear until hospital admission density plot

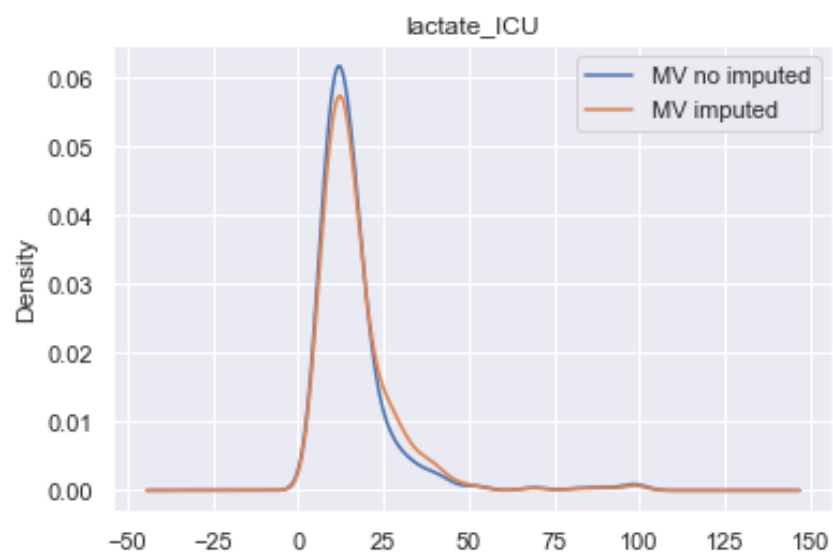


Figure 10.71: Lactate (ICU) density plot

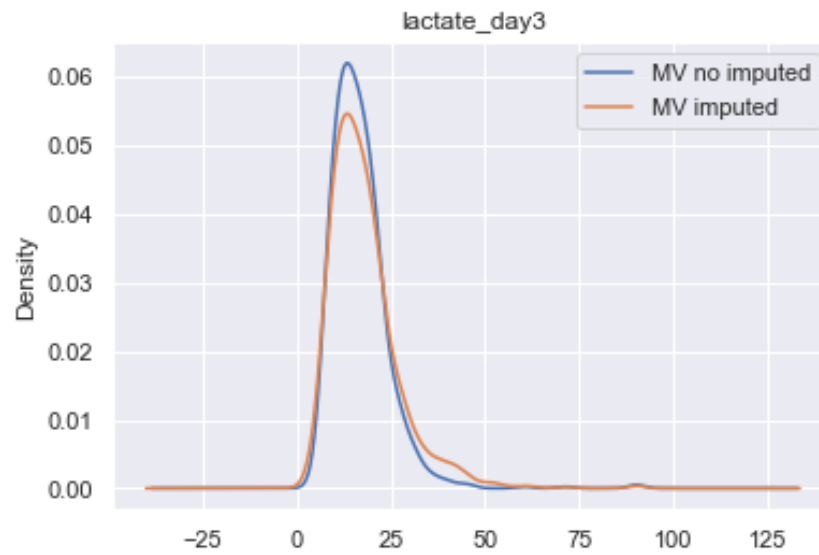


Figure 10.72: Lactate (3rd day ICU) density plot

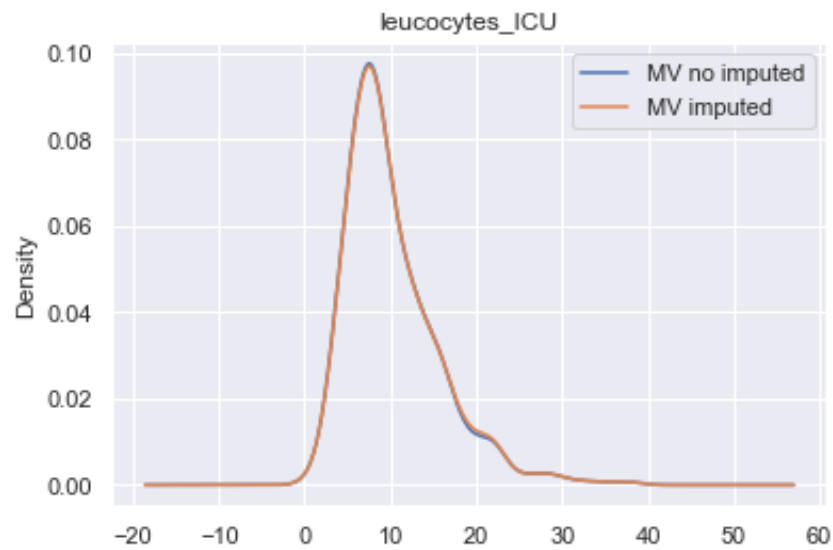


Figure 10.73: Leucocytes (ICU) density plot

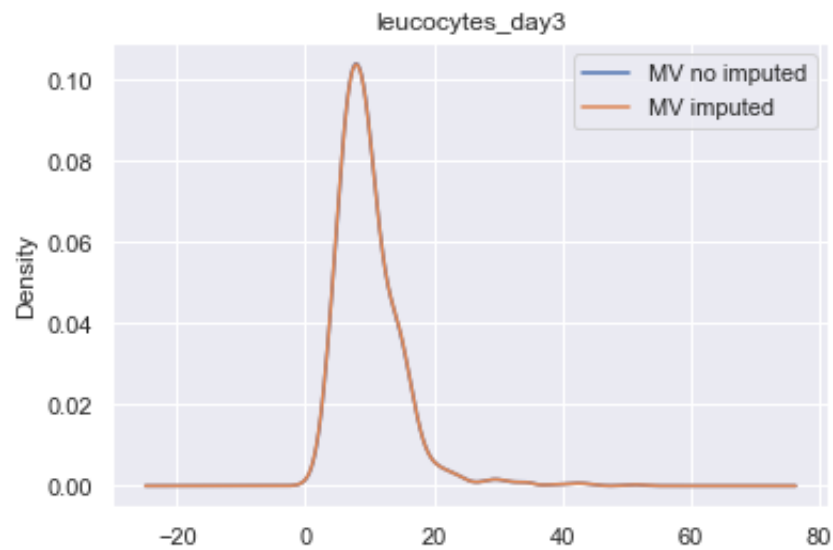


Figure 10.74: Leucocytes (3rd day ICU) density plot

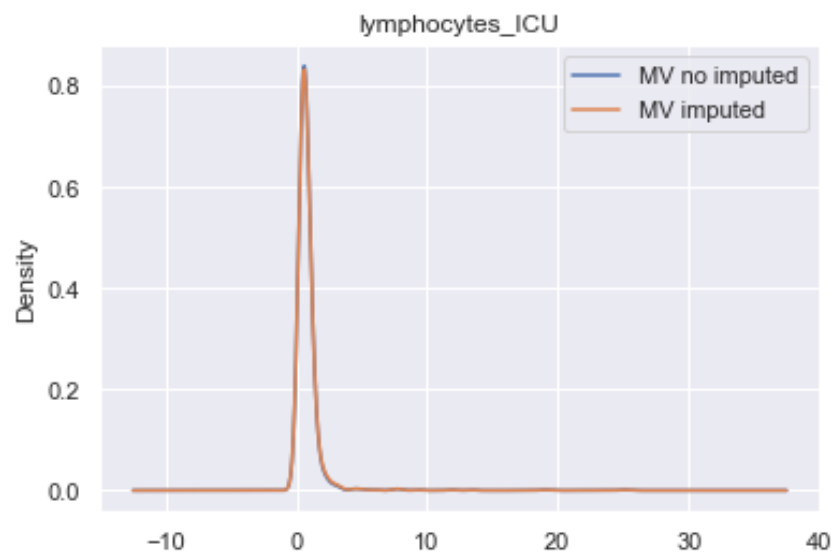


Figure 10.75: Lymphocytes (ICU) density plot

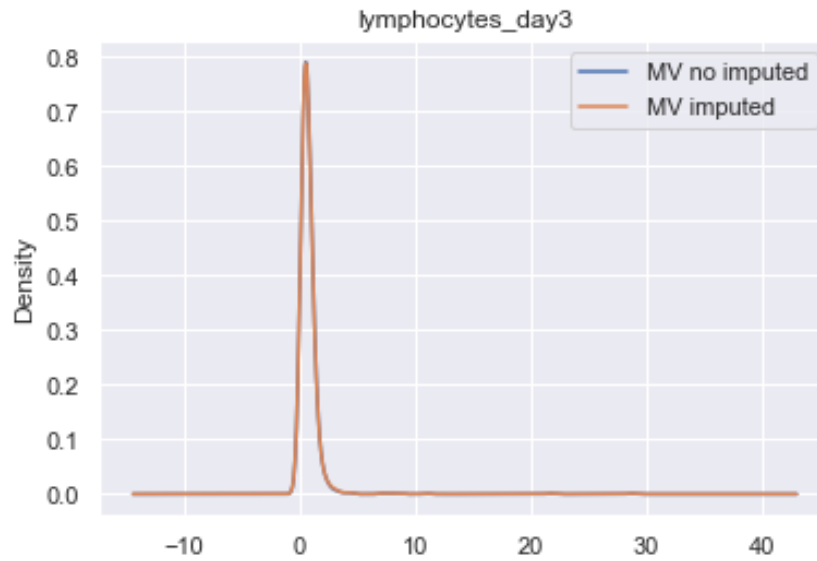


Figure 10.76: Lymphocytes (3rd day ICU) density plot

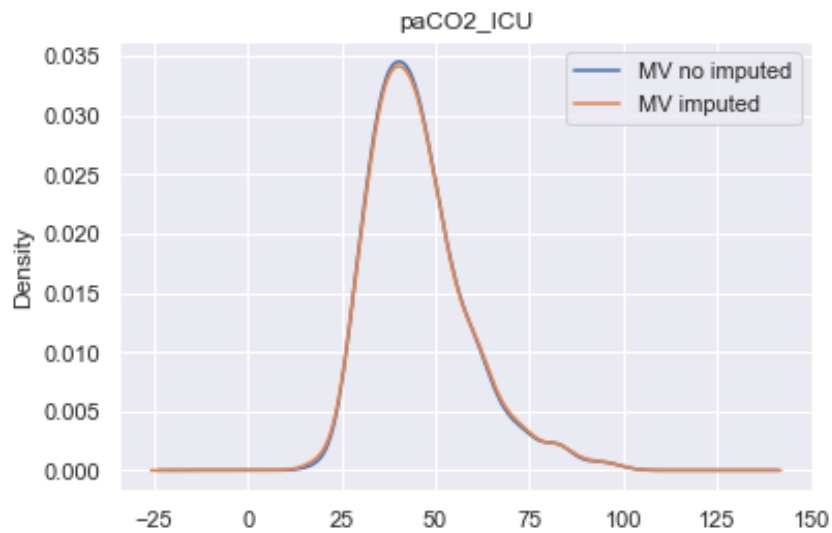


Figure 10.77: Carbon dioxide blood pressure (PaCO₂) (ICU) density plot

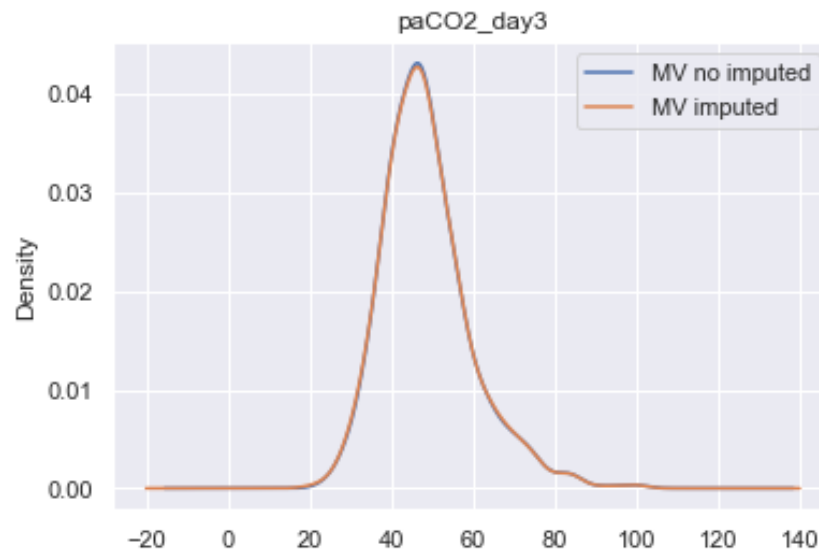


Figure 10.78: Carbon dioxide blood pressure (PaCO₂) (3rd day ICU) density plot

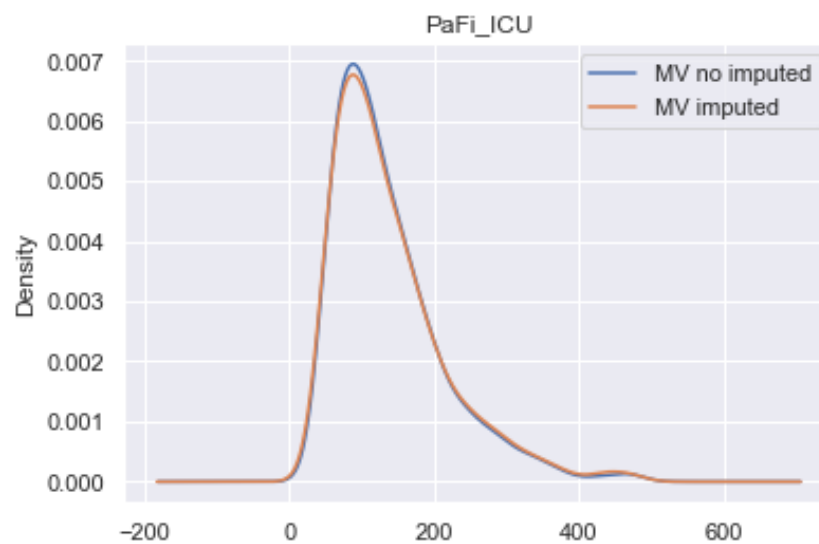


Figure 10.79: Platelet-aggregation factor inhibitor (PaFi) (ICU) density plot

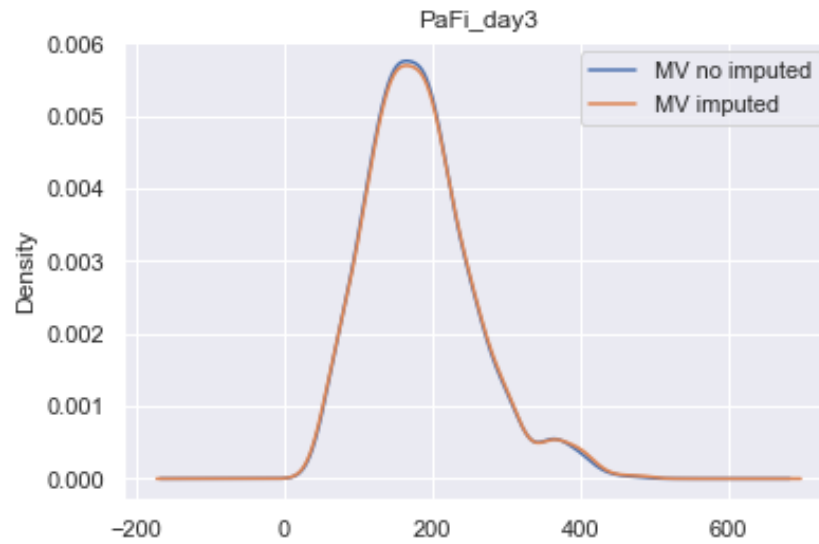


Figure 10.80: Platelet-aggregation factor inhibitor (PaFi) (3rd day ICU) density plot

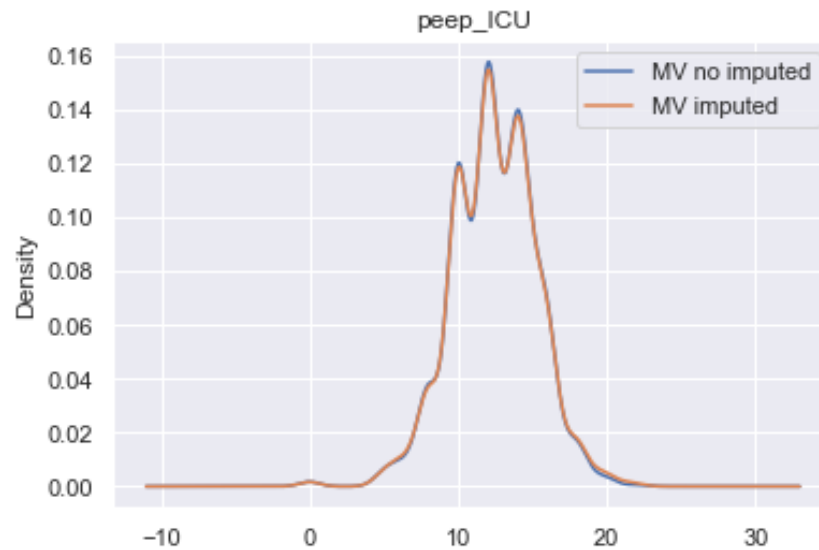


Figure 10.81: Positive pressure at the end of the expiration date (PEEP) (ICU) density plot

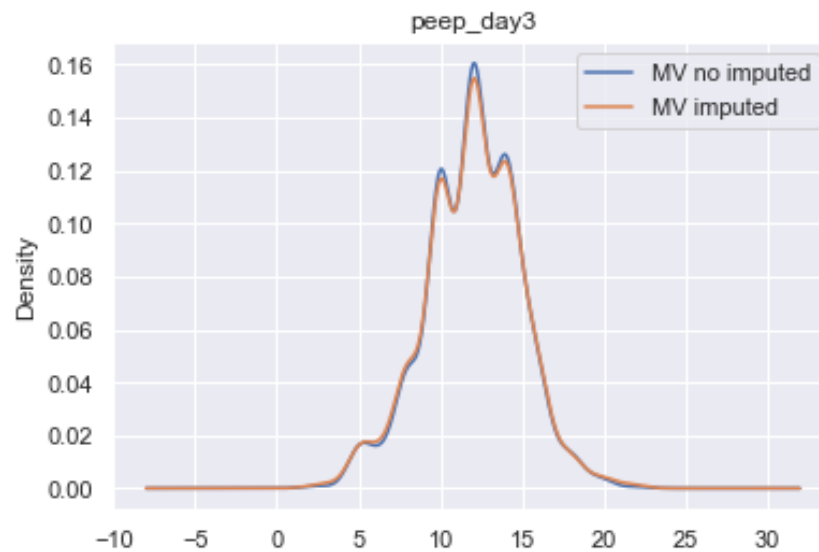


Figure 10.82: Positive pressure at the end of the expiration date (PEEP) (3rd day ICU) density plot

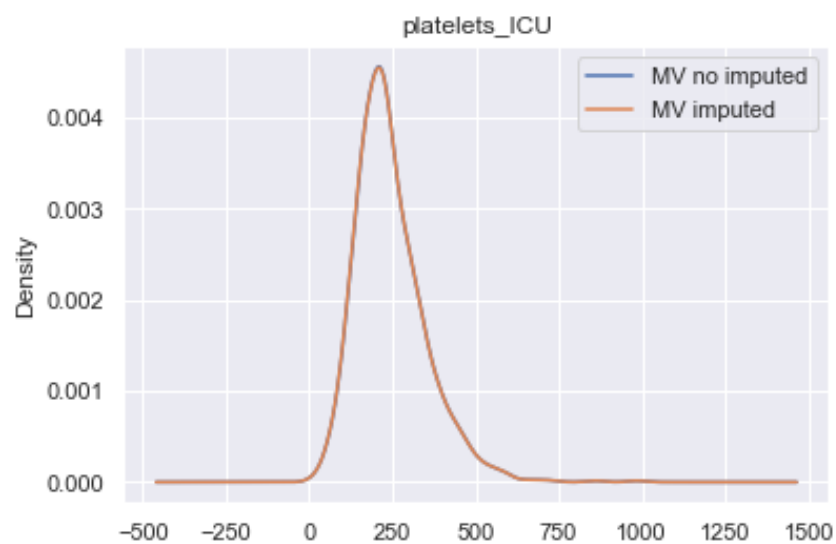


Figure 10.83: Platelets (ICU) density plot

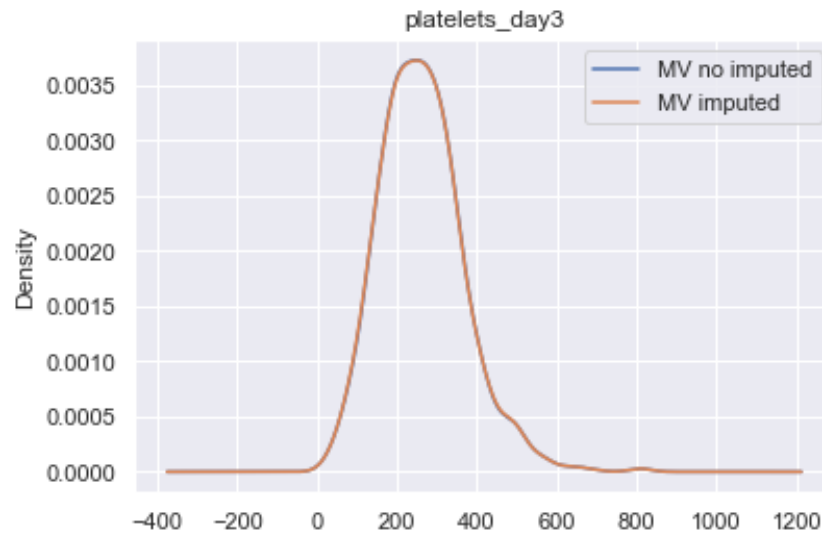


Figure 10.84: platelets (3rd day ICU) density plot

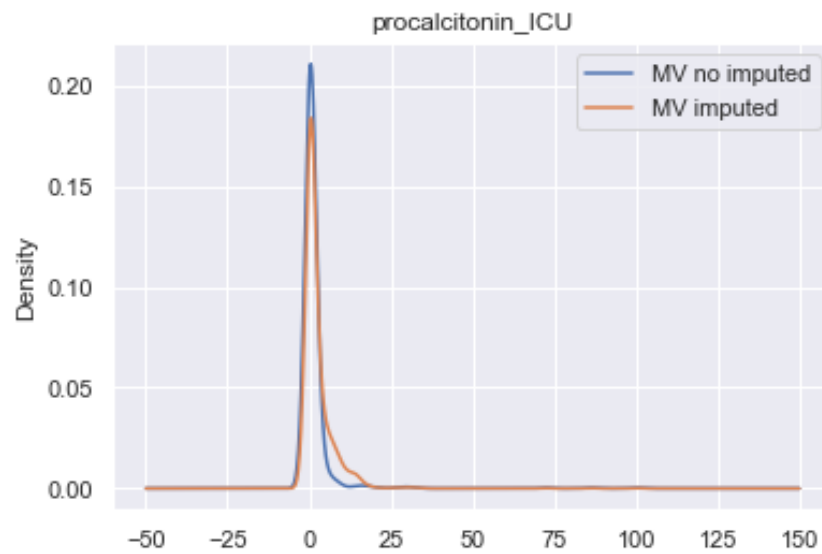


Figure 10.85: Procalcitonin (ICU) density plot

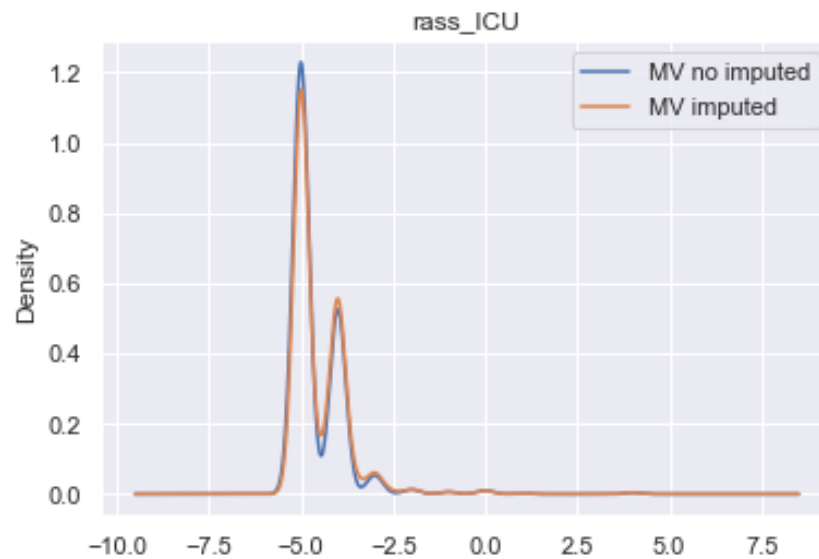


Figure 10.86: Richmond Agitation Sedation Scale (RASS)(ICU) density plot

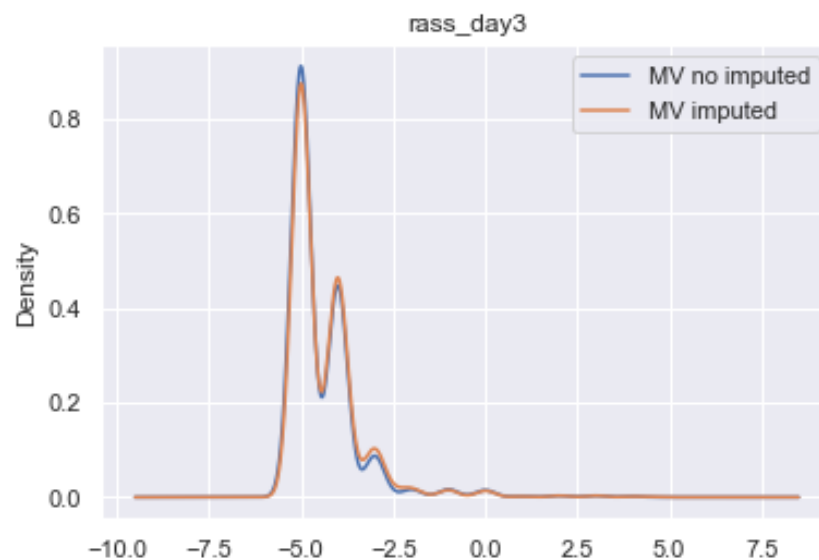


Figure 10.87: Richmond Agitation Sedation Scale (RASS) (3rd day ICU) density plot

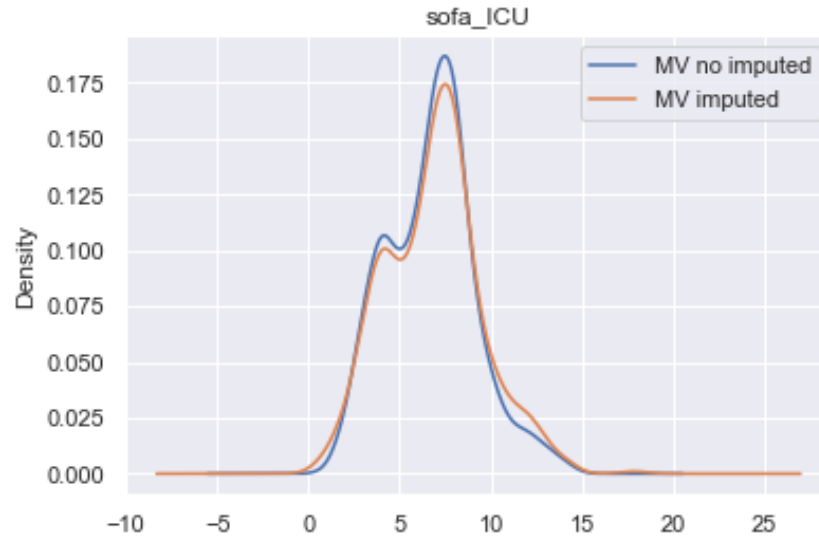


Figure 10.88: Sequential Organ Failure Assessment score (SOFA) (ICU) density plot

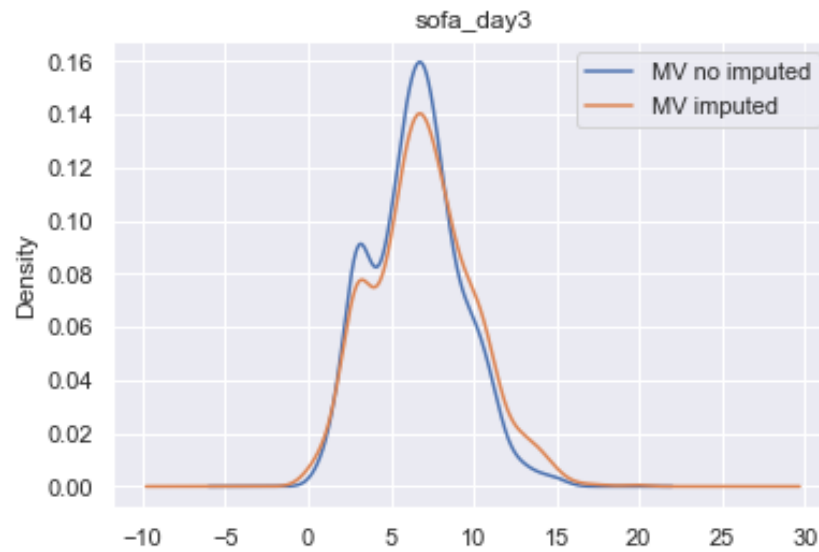


Figure 10.89: Sequential Organ Failure Assessment score (SOFA) (3rd day ICU) density plot

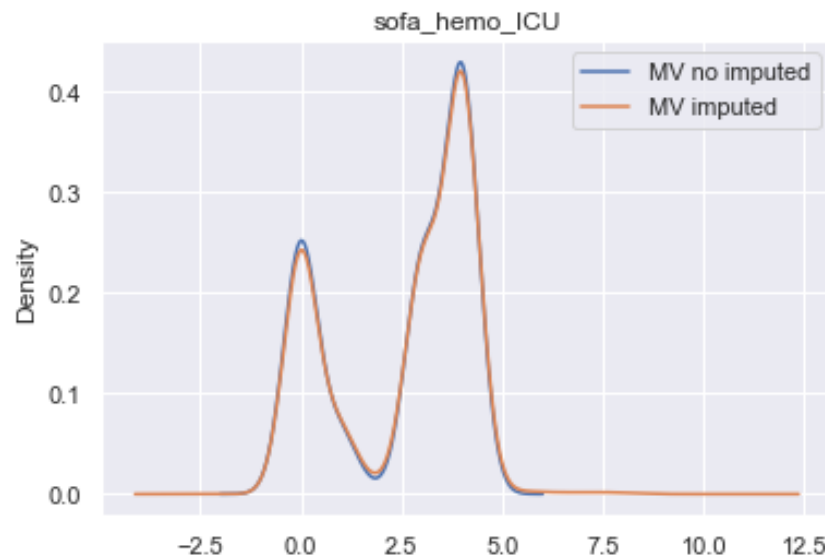


Figure 10.90: Hemodynamic Sequential Organ Failure Assessment score (hemodynamic SOFA) (ICU) density plot

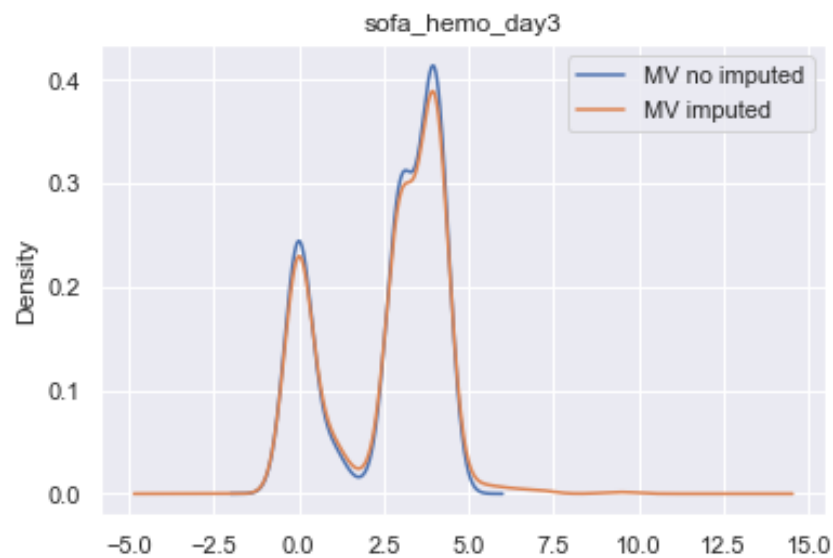


Figure 10.91: Hemodynamic Sequential Organ Failure Assessment score (hemodynamic SOFA) (3rd day ICU) density plot

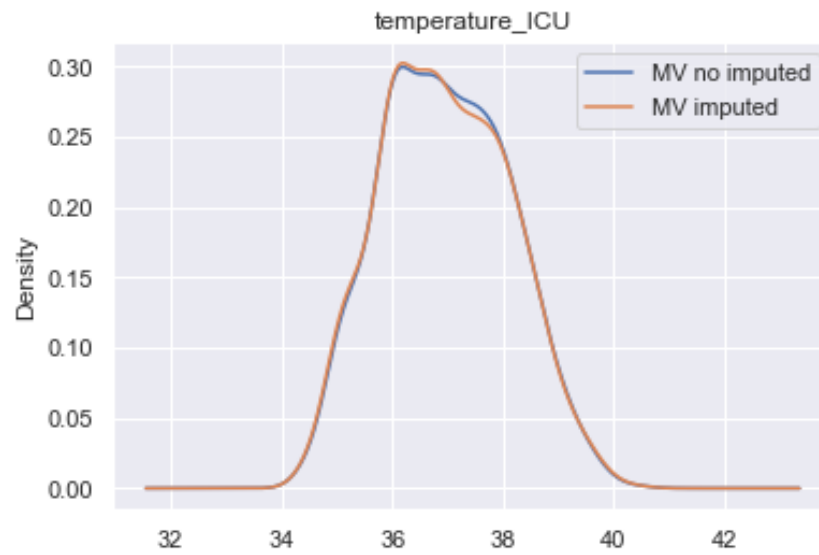


Figure 10.92: Temperature (ICU) density plot

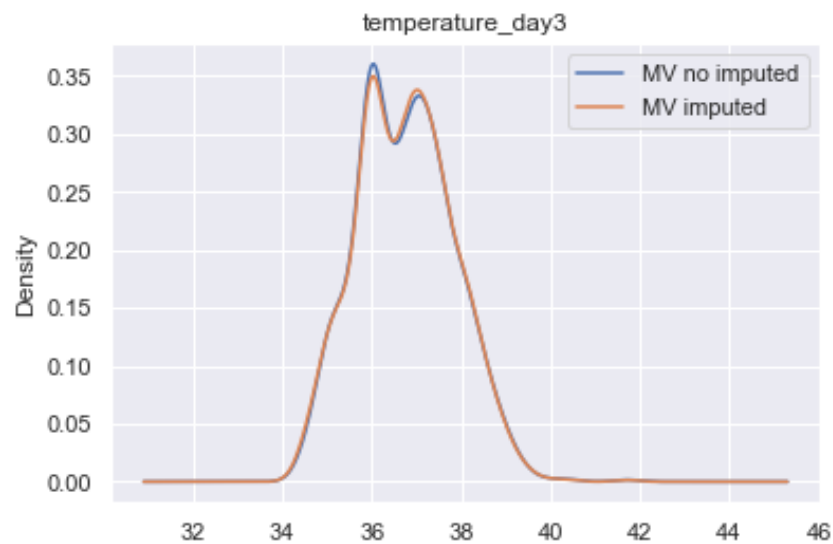


Figure 10.93: Temperature (3rd day ICU) density plot

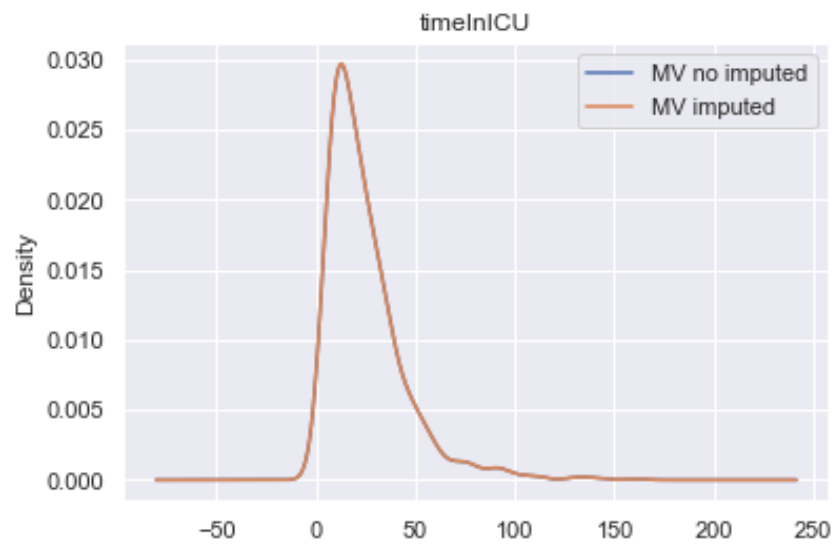


Figure 10.94: Time in ICU density plot

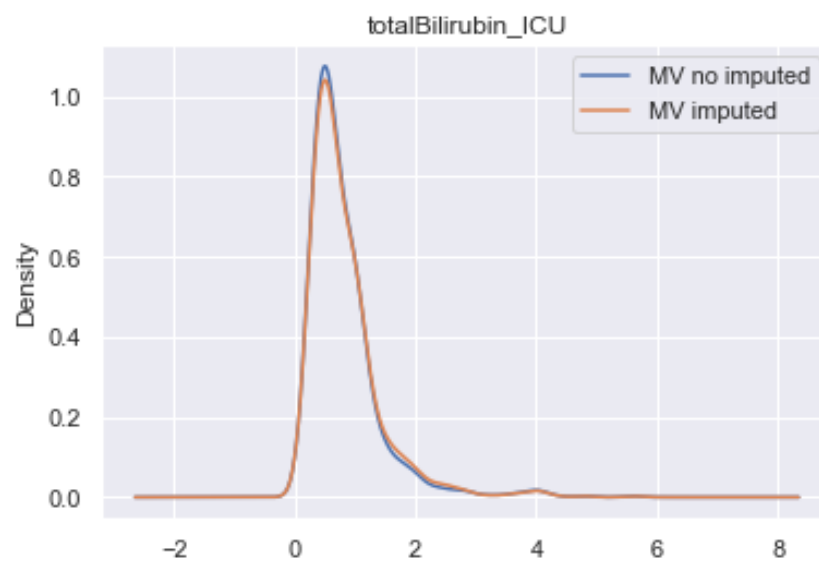


Figure 10.95: Total bilirubin (ICU) density plot

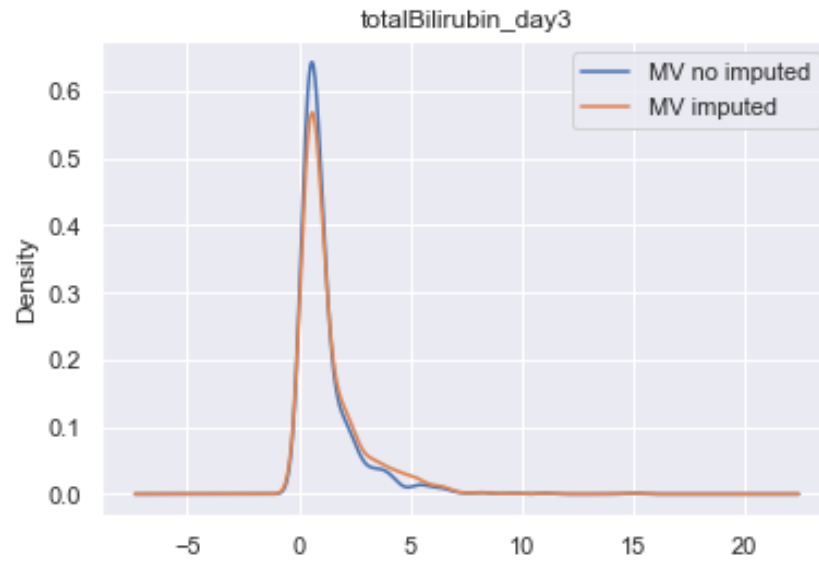


Figure 10.96: Total bilirubin (3rd day ICU) density plot

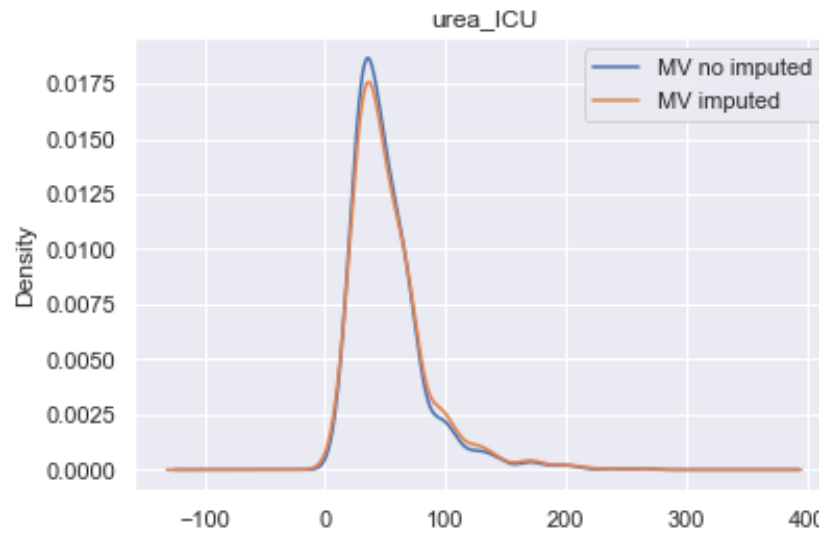


Figure 10.97: Urea (ICU) density plot

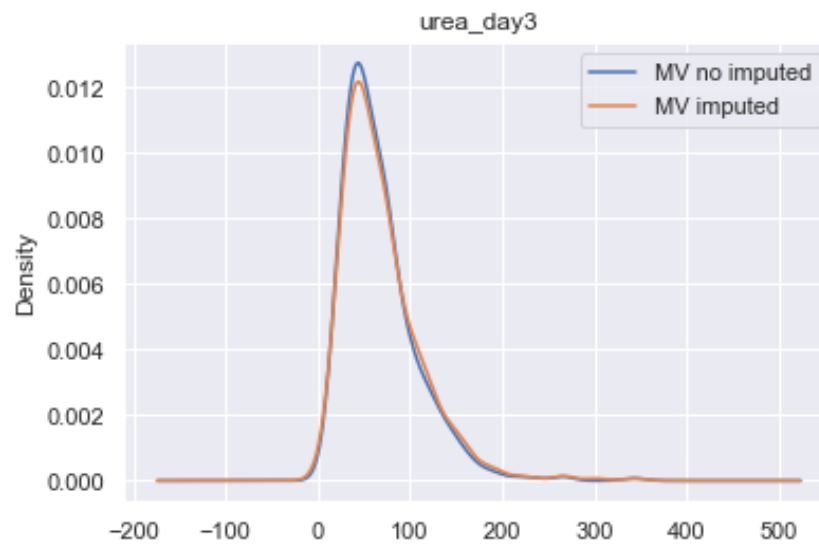


Figure 10.98: Urea (3rd day ICU) density plot

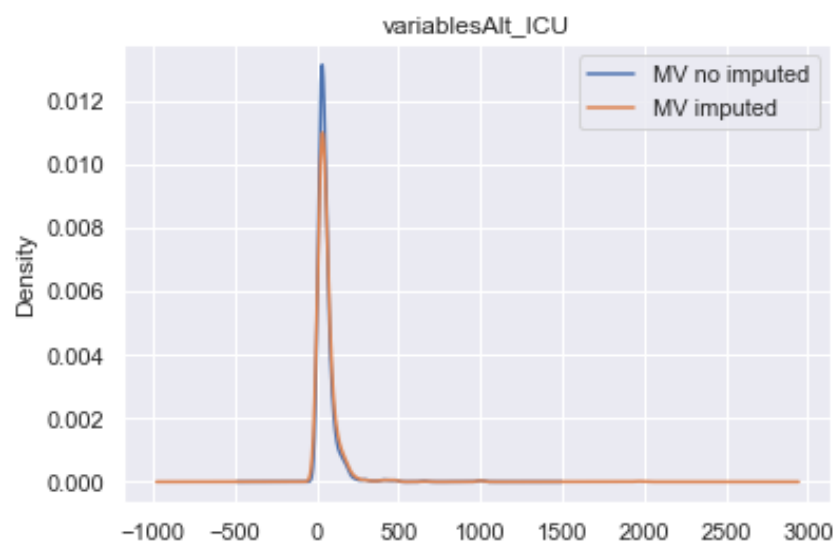


Figure 10.99: Alanine transaminase (ALT) (ICU) density plot

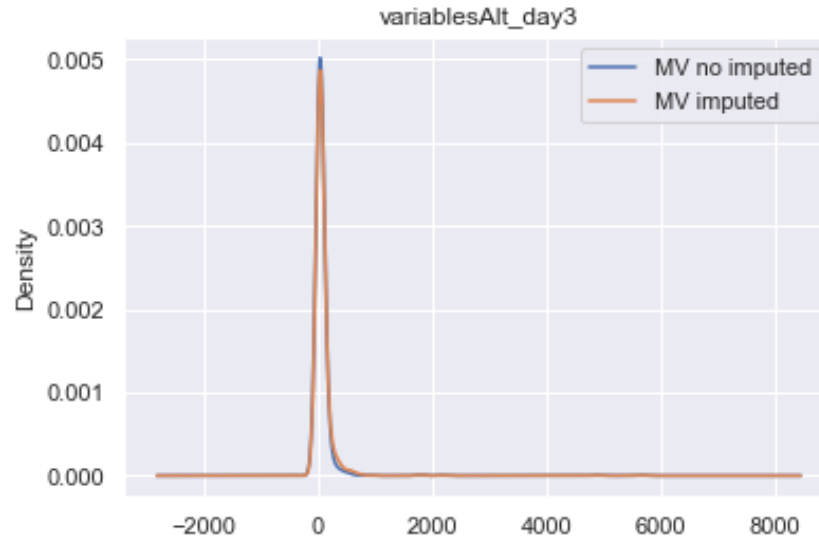


Figure 10.100: Alanine transaminase (ALT) (3rd day ICU) density plot

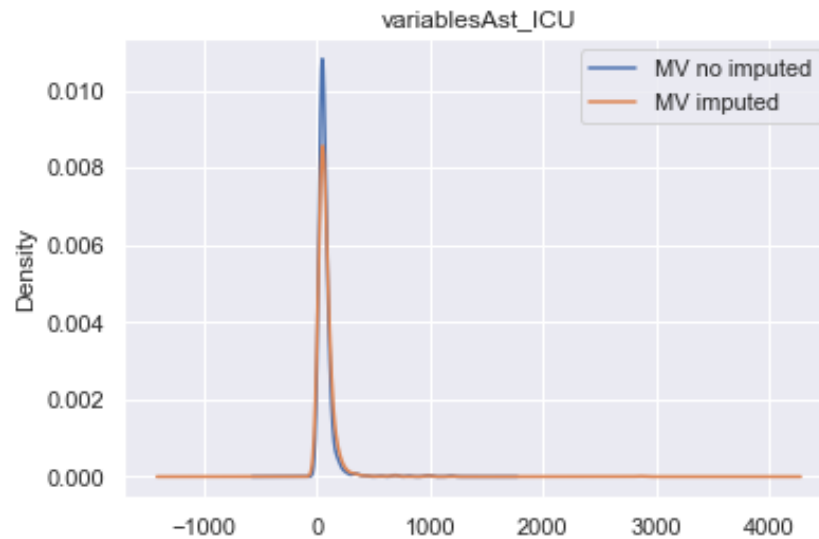


Figure 10.101: Aspartate transaminase (AST) (ICU) density plot

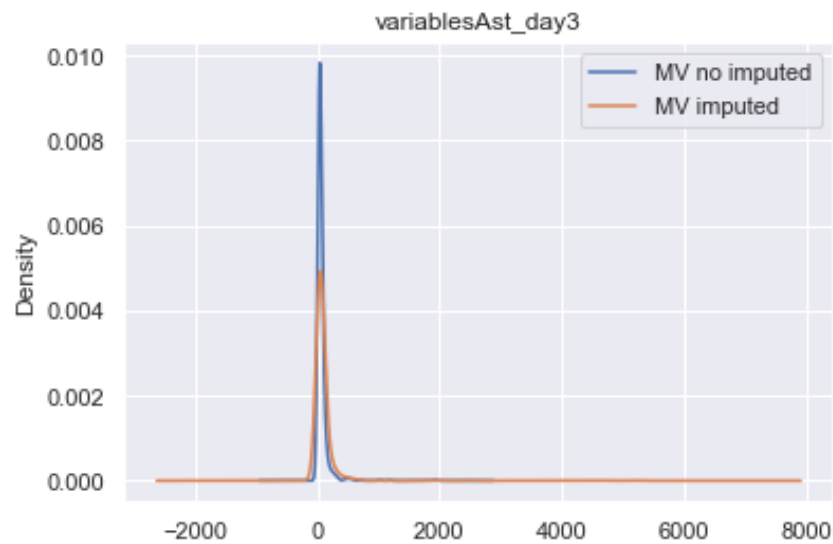


Figure 10.102: Aspartate transaminase (AST) (3rd day ICU) density plot

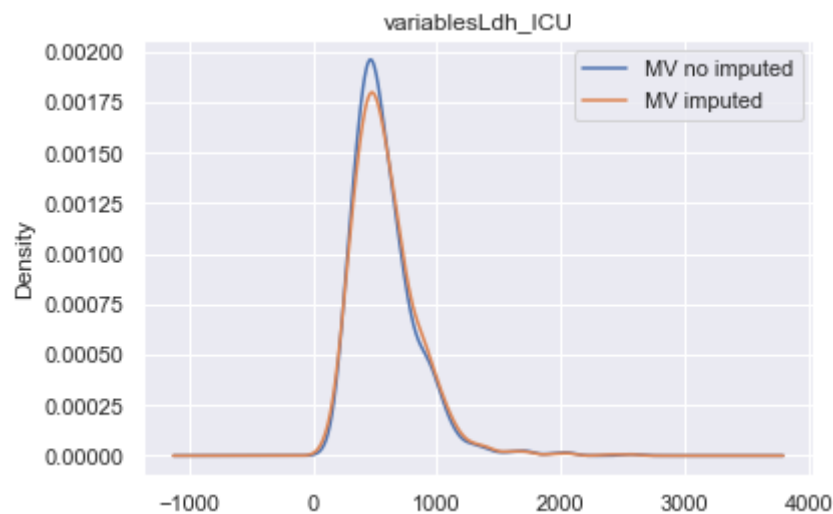


Figure 10.103: Lactate dehydrogenase (LDH) (ICU) density plot

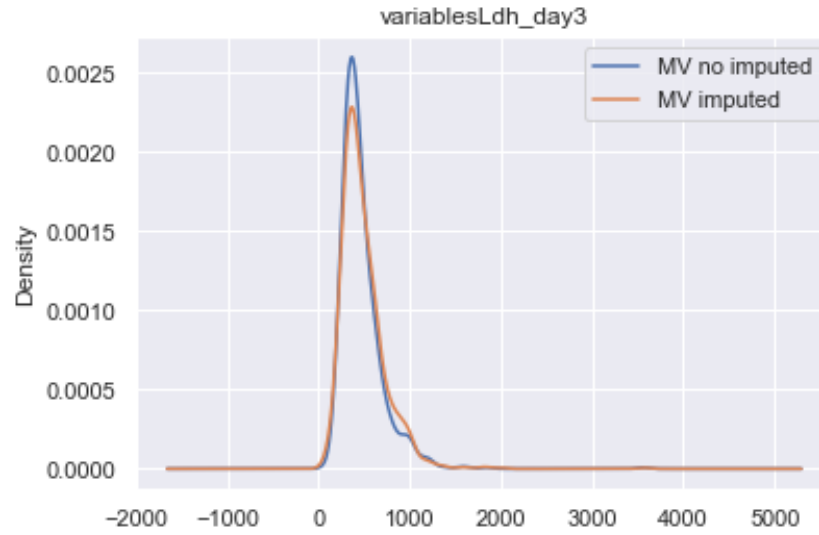


Figure 10.104: Lactate dehydrogenase (LDH) (3rd day ICU) density plot

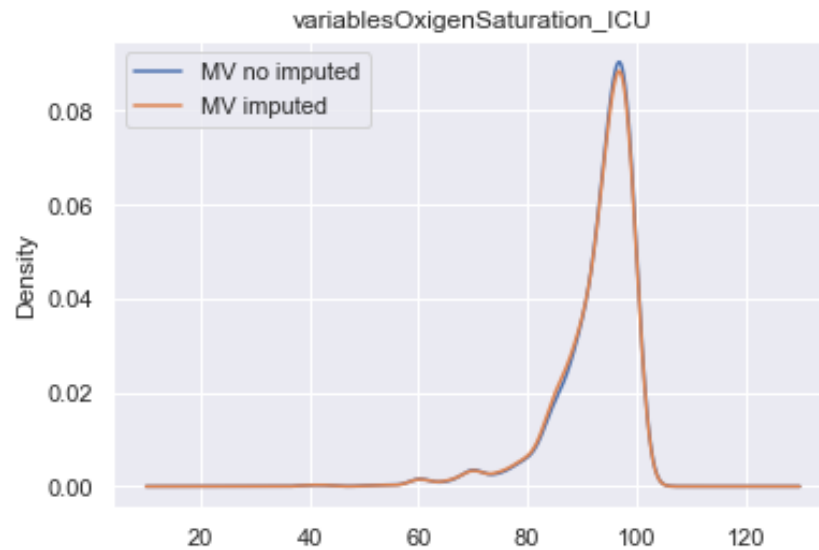


Figure 10.105: Oxygen saturation (ICU) density plot

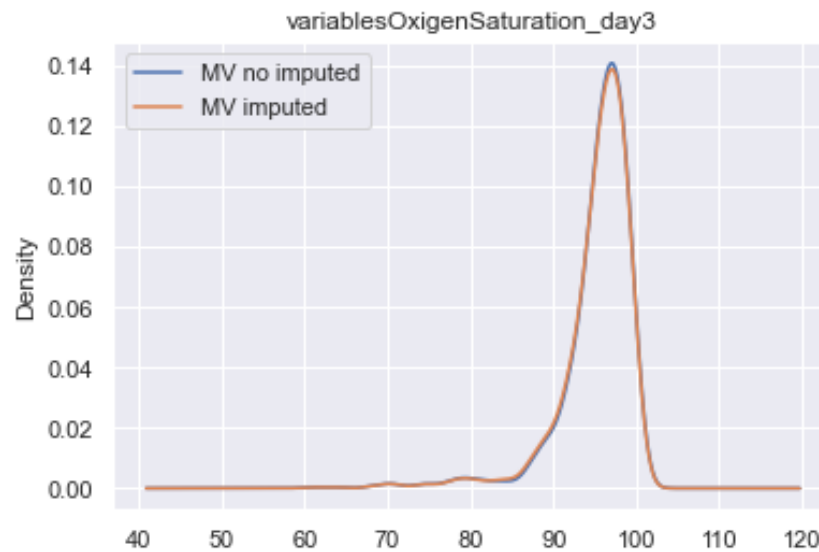


Figure 10.106: Oxygen saturation (3rd day ICU) density plot

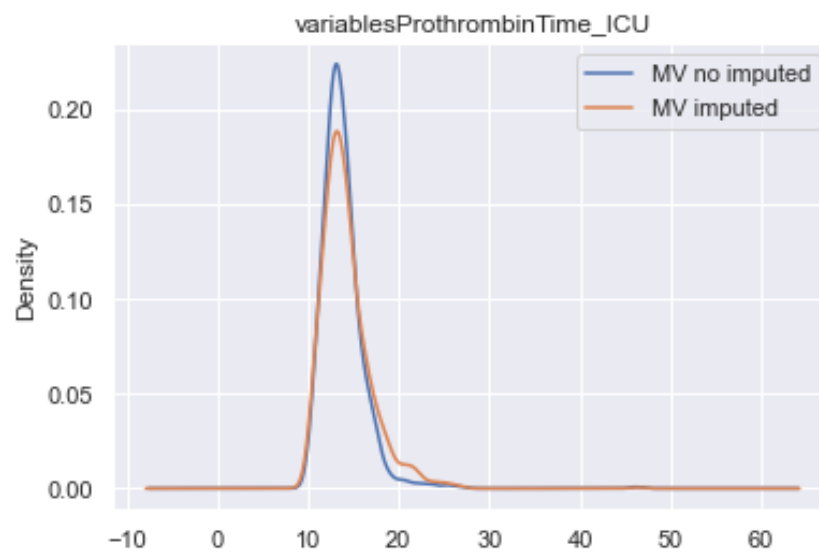


Figure 10.107: Prothrombin time (PT) (ICU) density plot

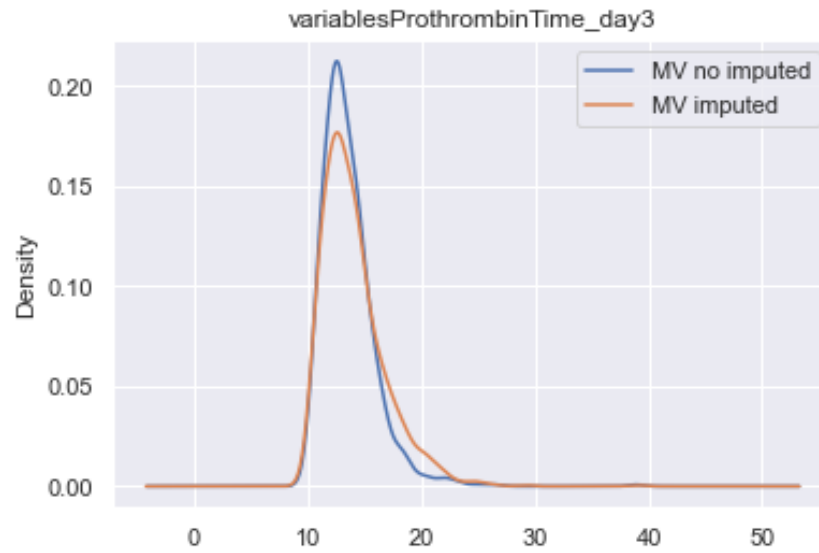


Figure 10.108: Prothrombin time (PT) (3rd day ICU) density plot

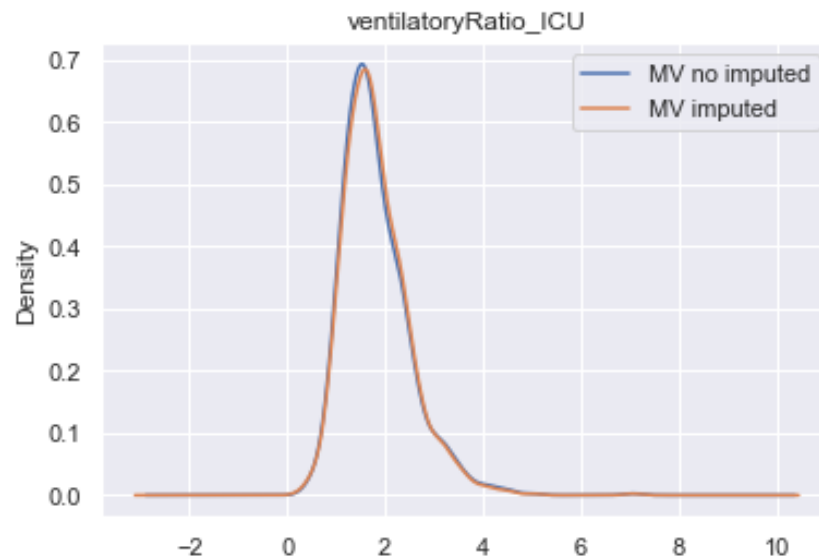


Figure 10.109: Ventilatory ratio (ICU) density plot

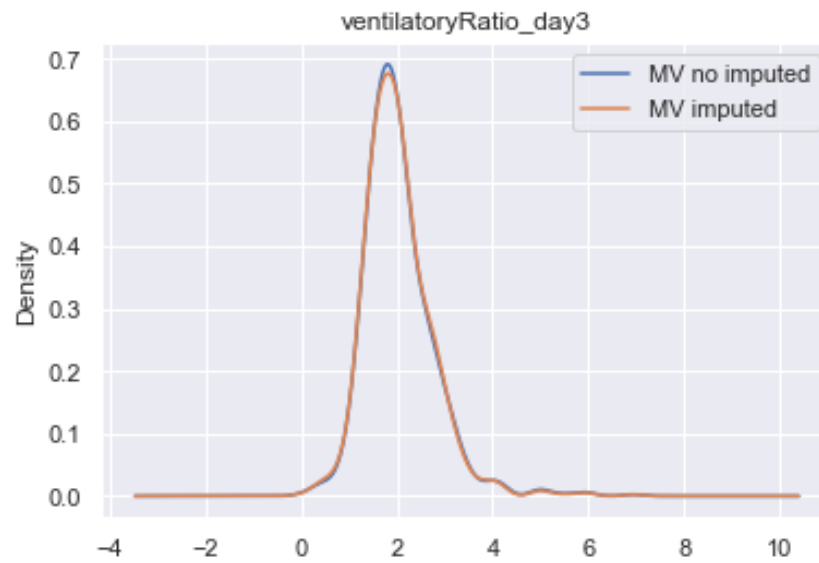


Figure 10.110: Ventilatory ratio (3rd day ICU) density plot

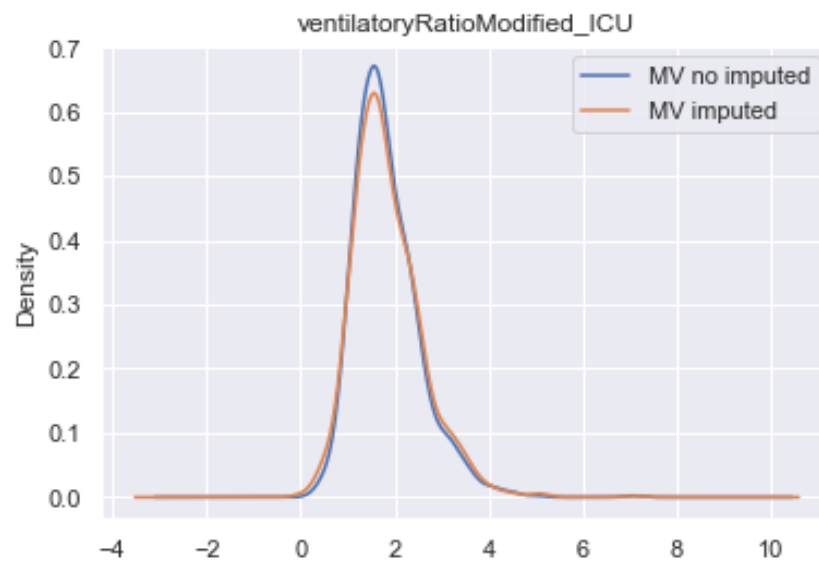


Figure 10.111: Ventilatory ratio *modified* (ICU) density plot

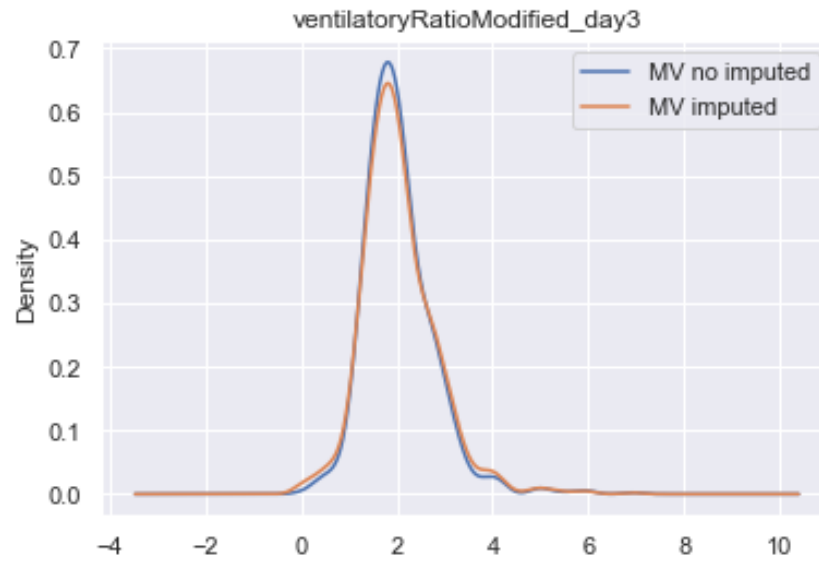


Figure 10.112: Ventilatory ratio *modified* (3rd day ICU) density plot

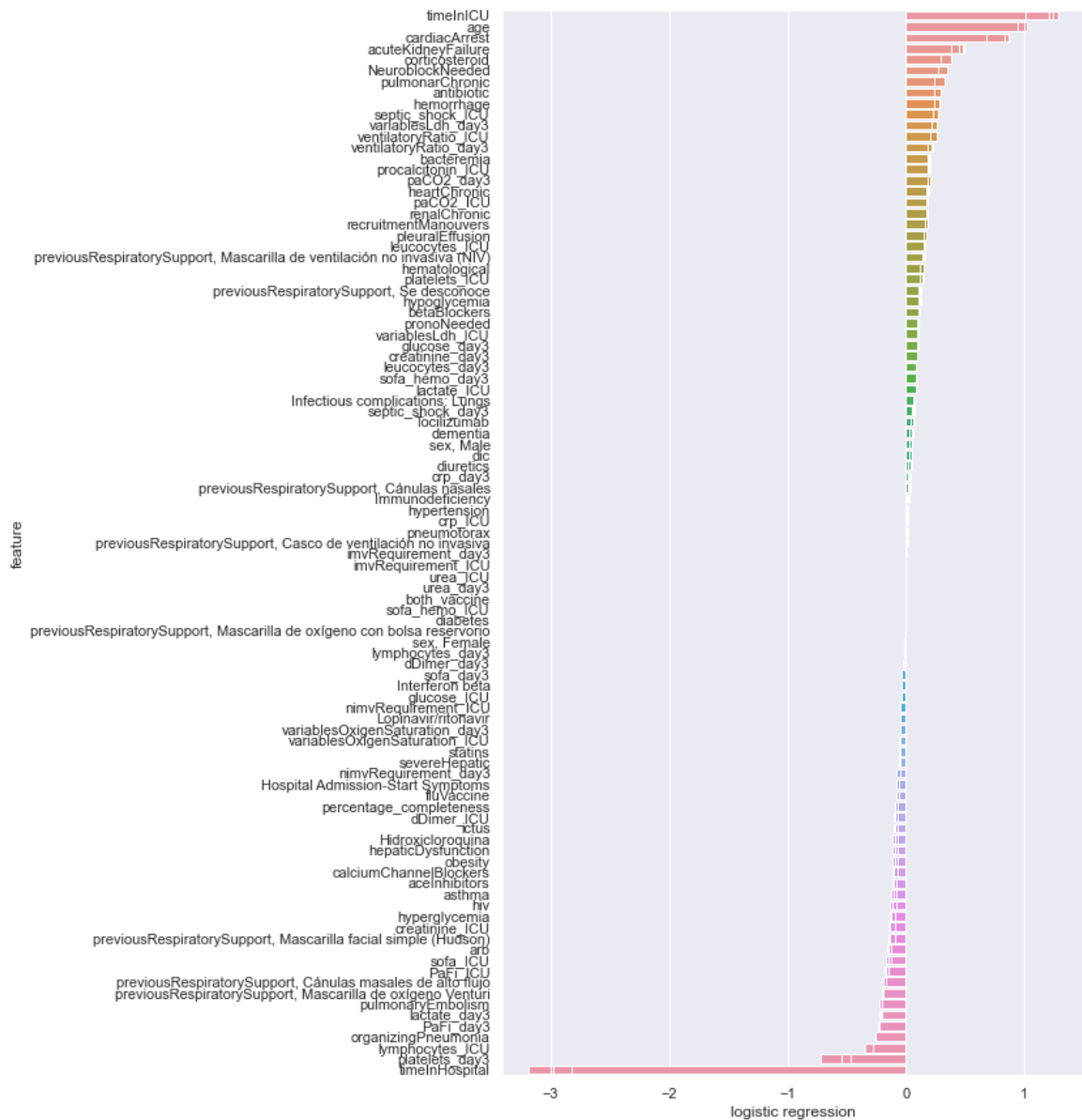


Figure 10.113: Logistic regression feature selection

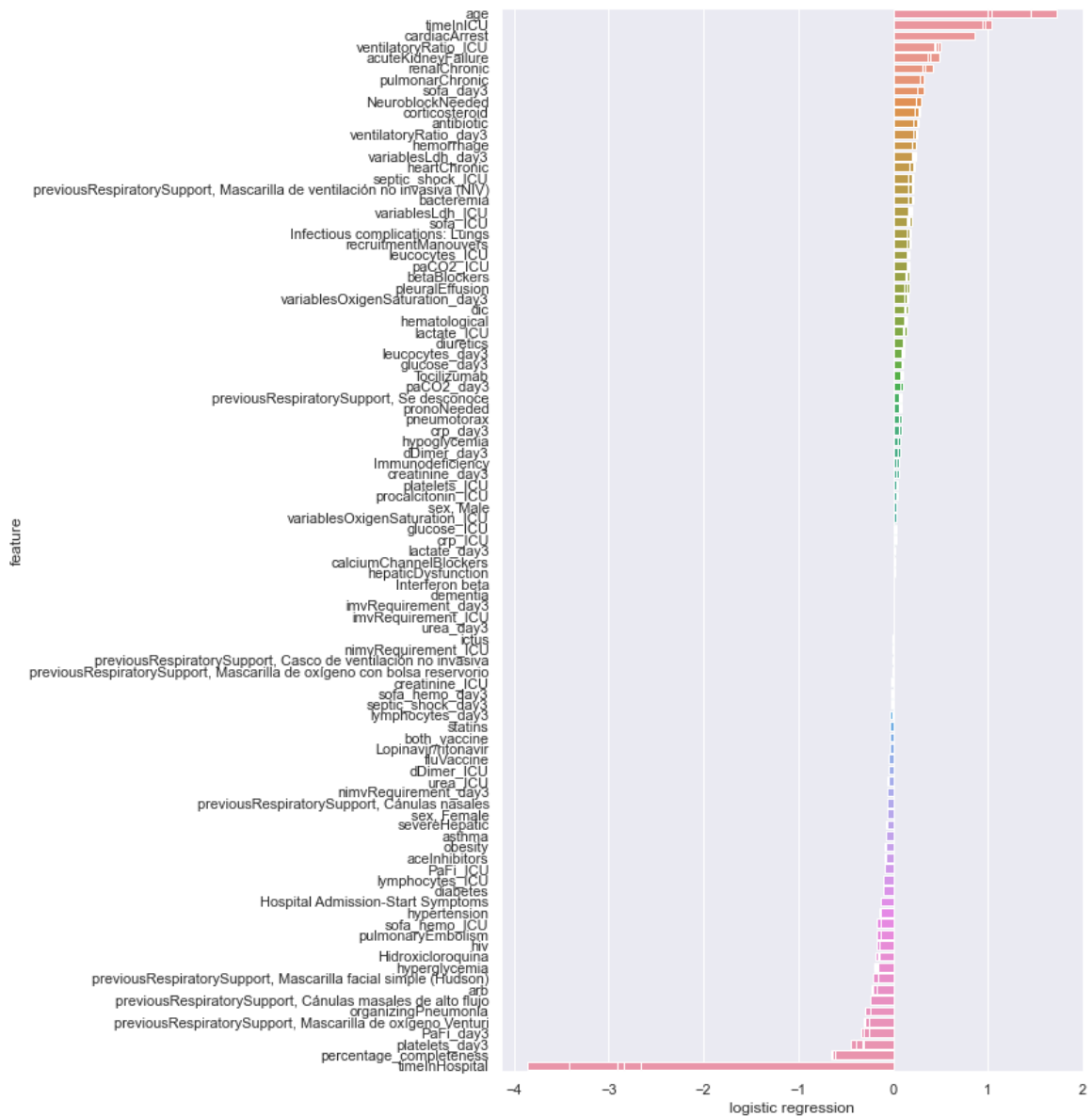


Figure 10.114: Logistic regression feature selection with imputed features

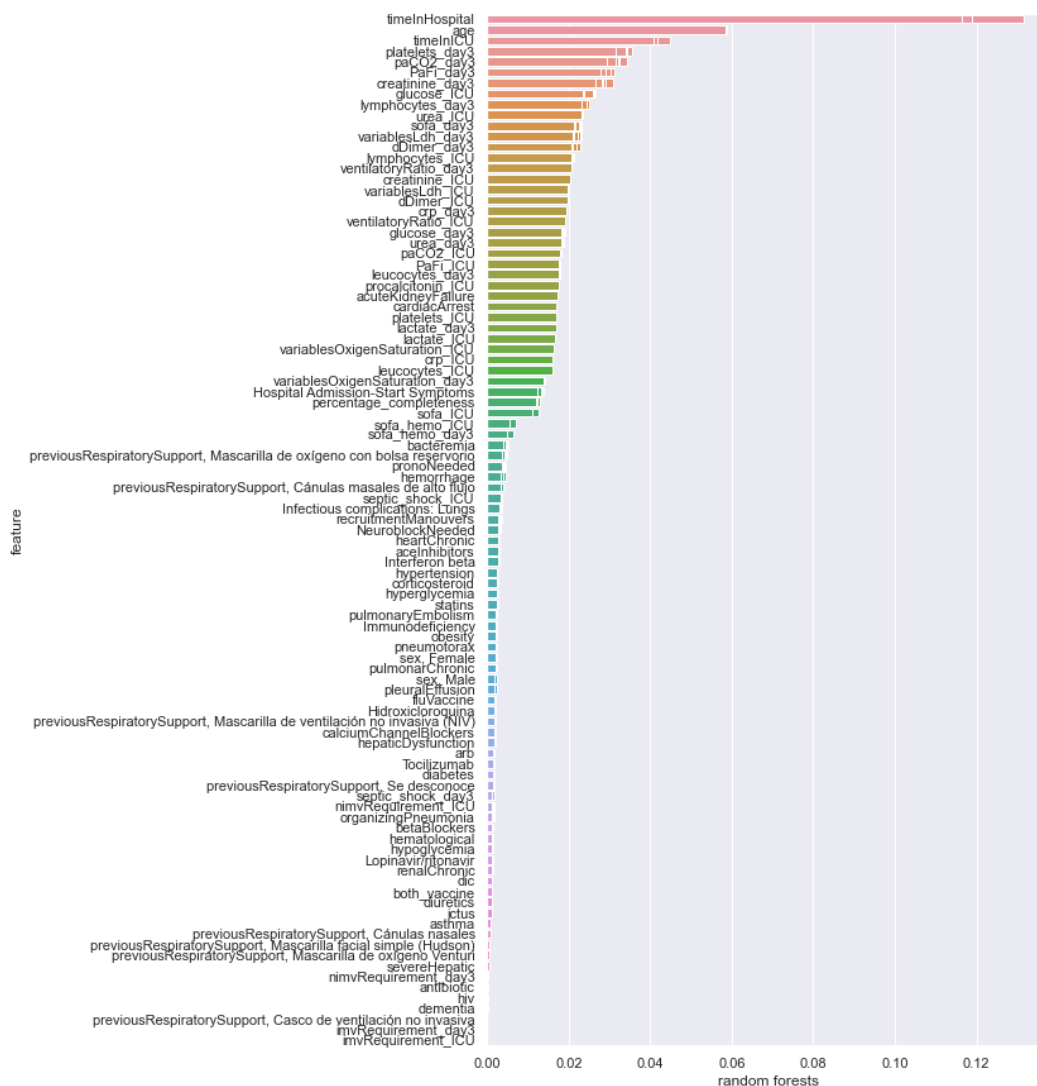


Figure 10.115: Random forest feature selection with imputed features

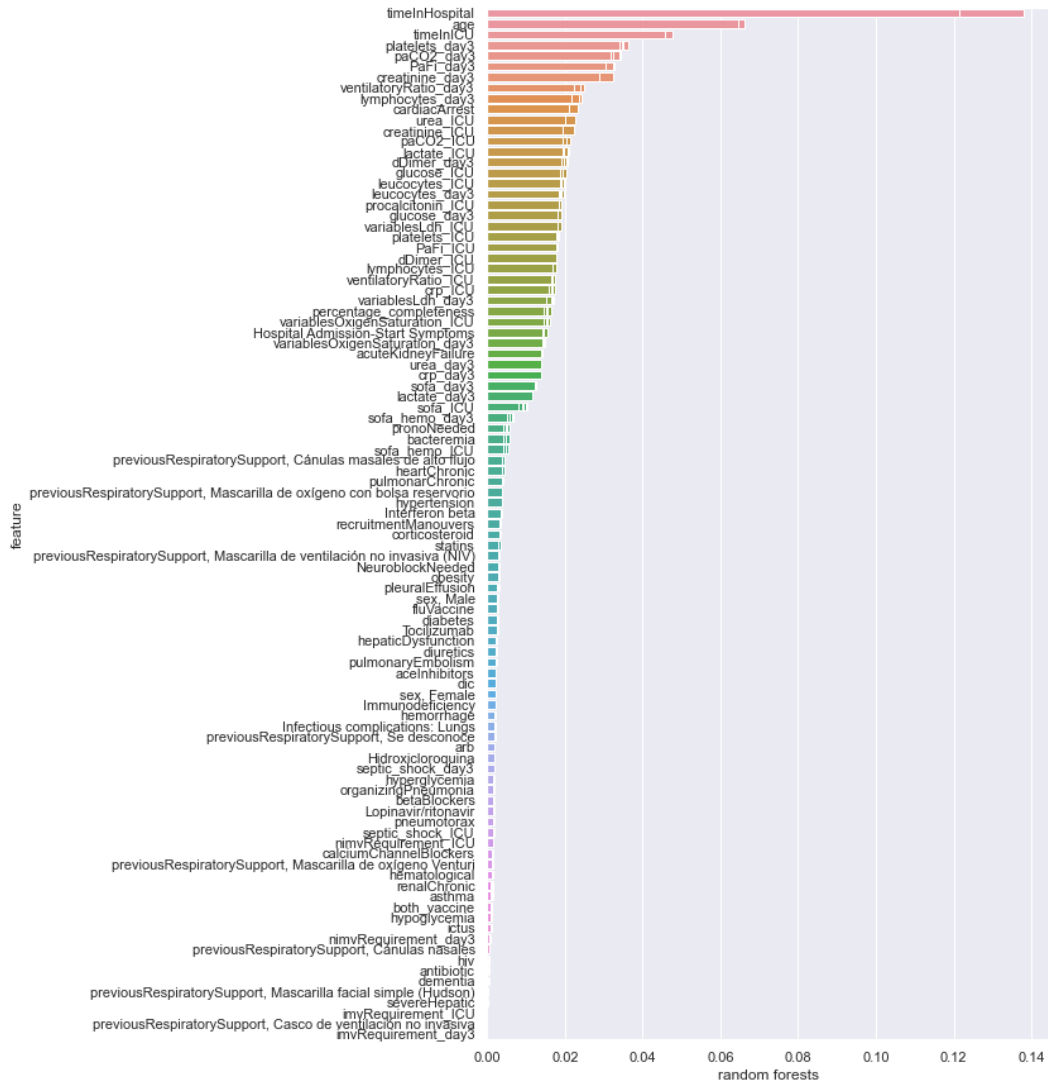


Figure 10.116: Random forest feature selection

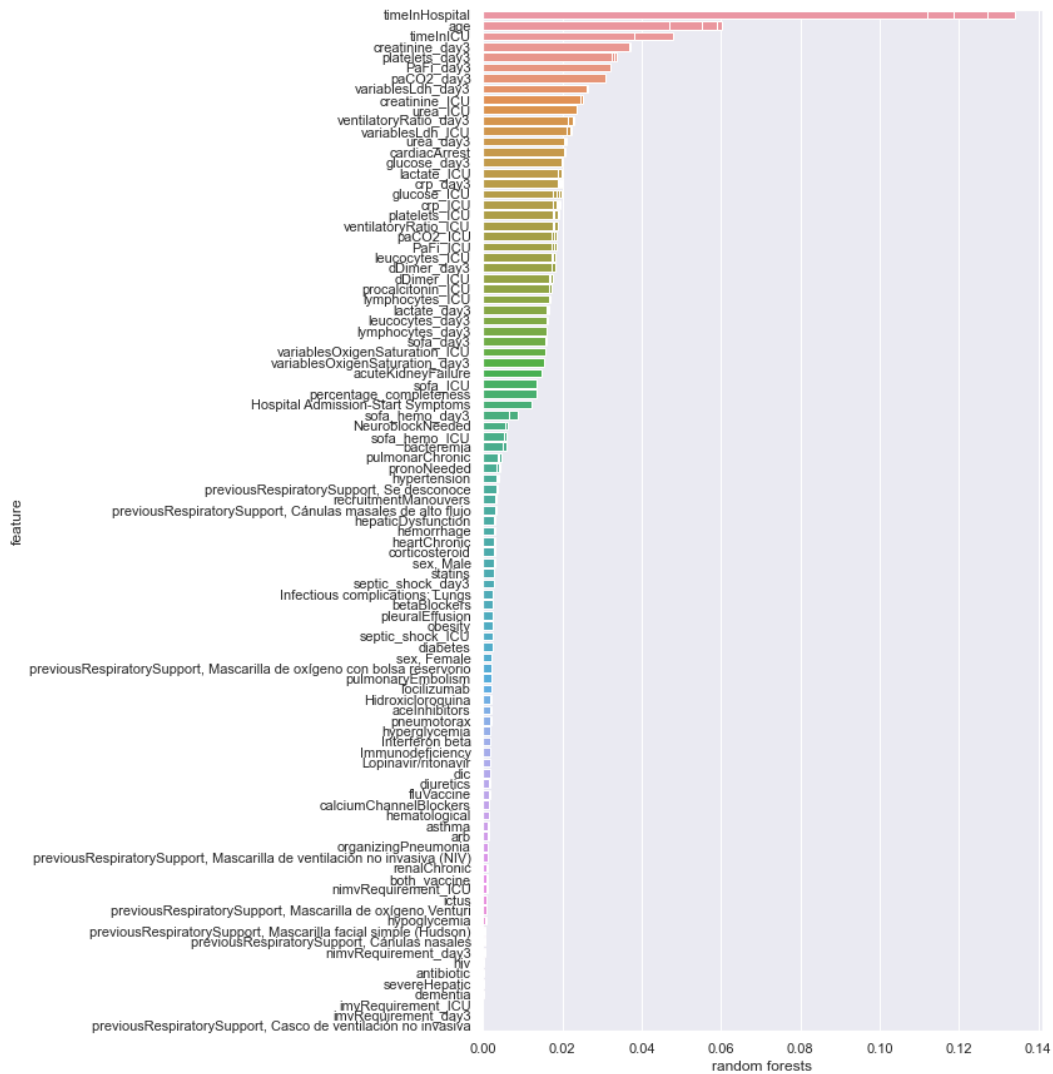


Figure 10.117: Recursive feature elimination with random forest with imputed features

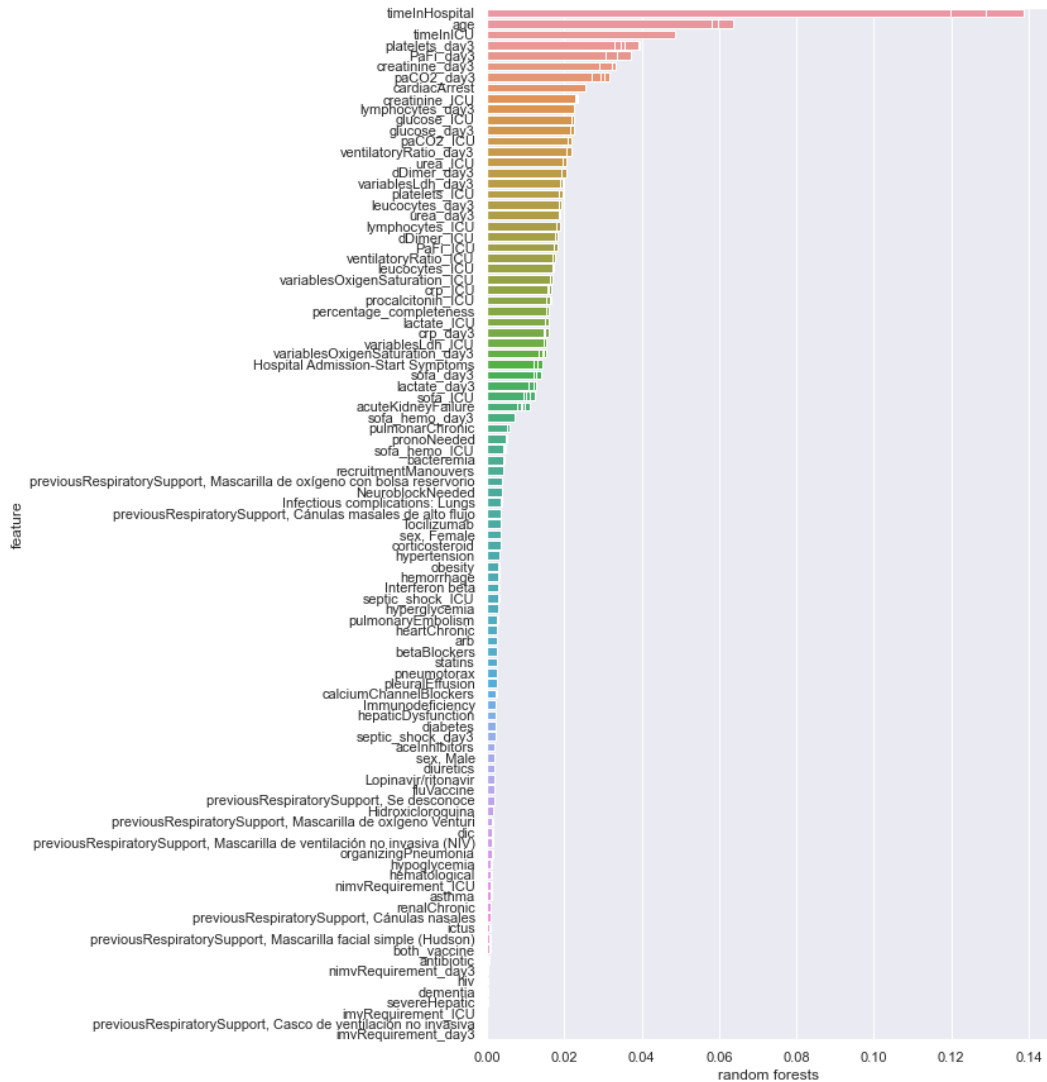


Figure 10.118: Recursive feature elimination with random forest

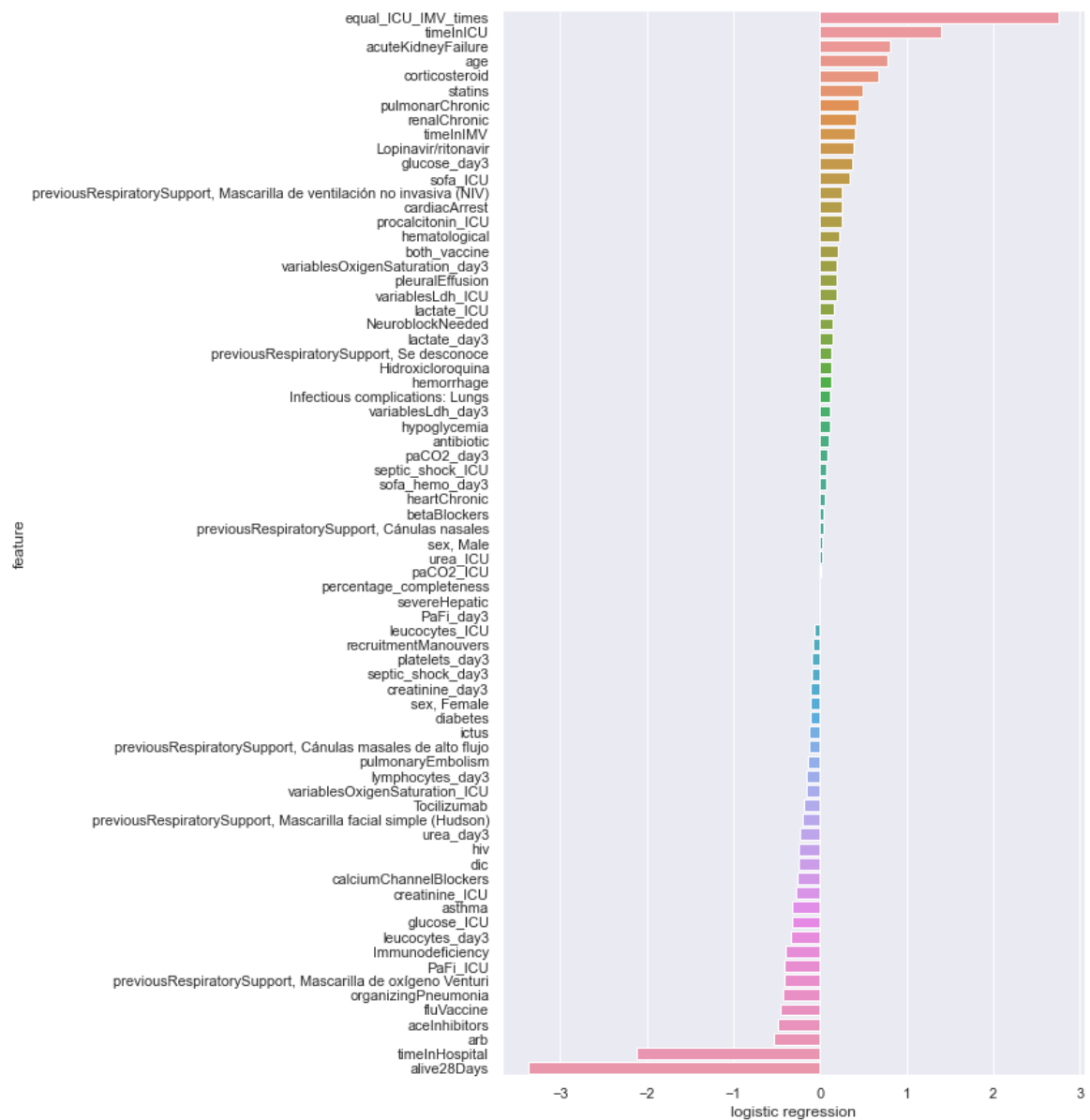


Figure 10.119: Recursive feature elimination with logistic regression with imputed features

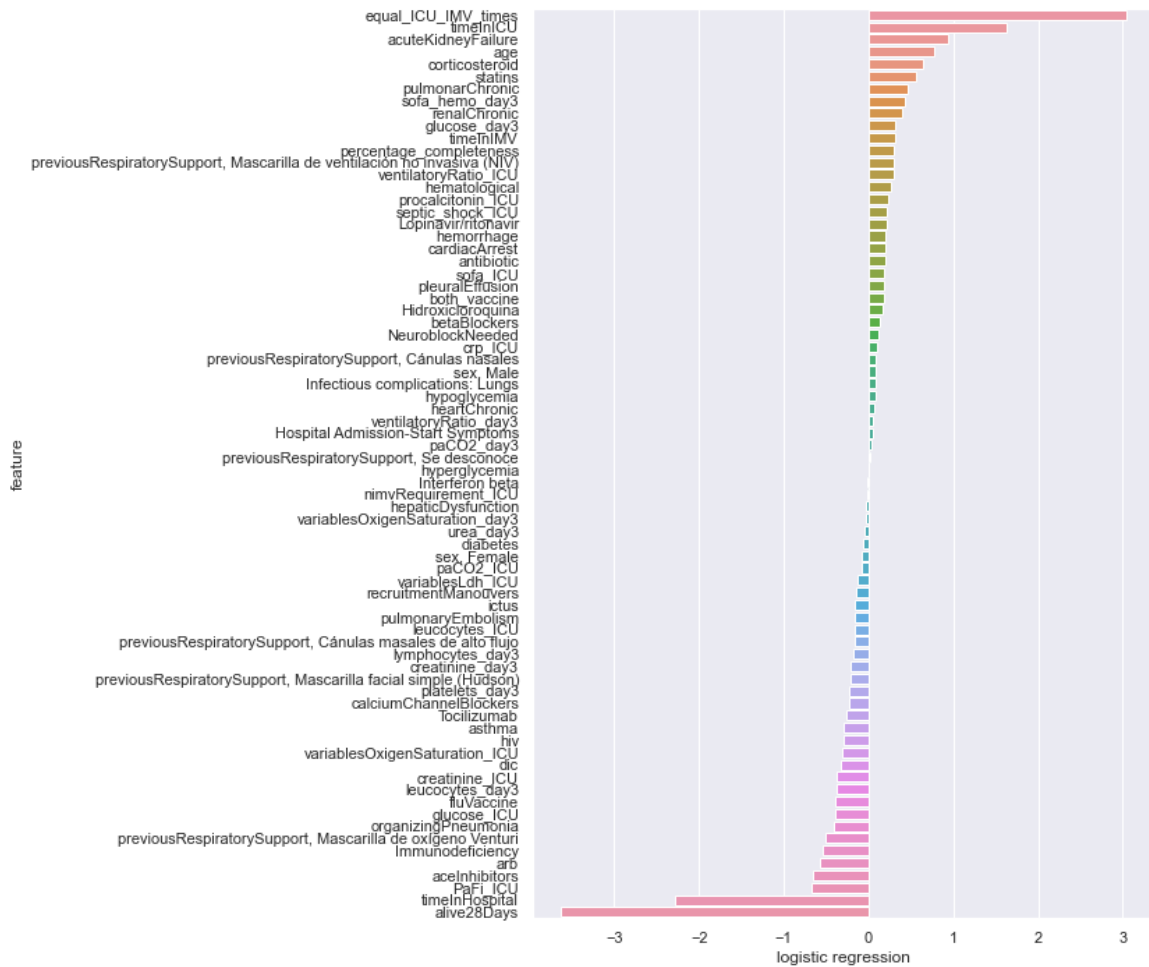


Figure 10.120: Recursive feature elimination with logistic regression

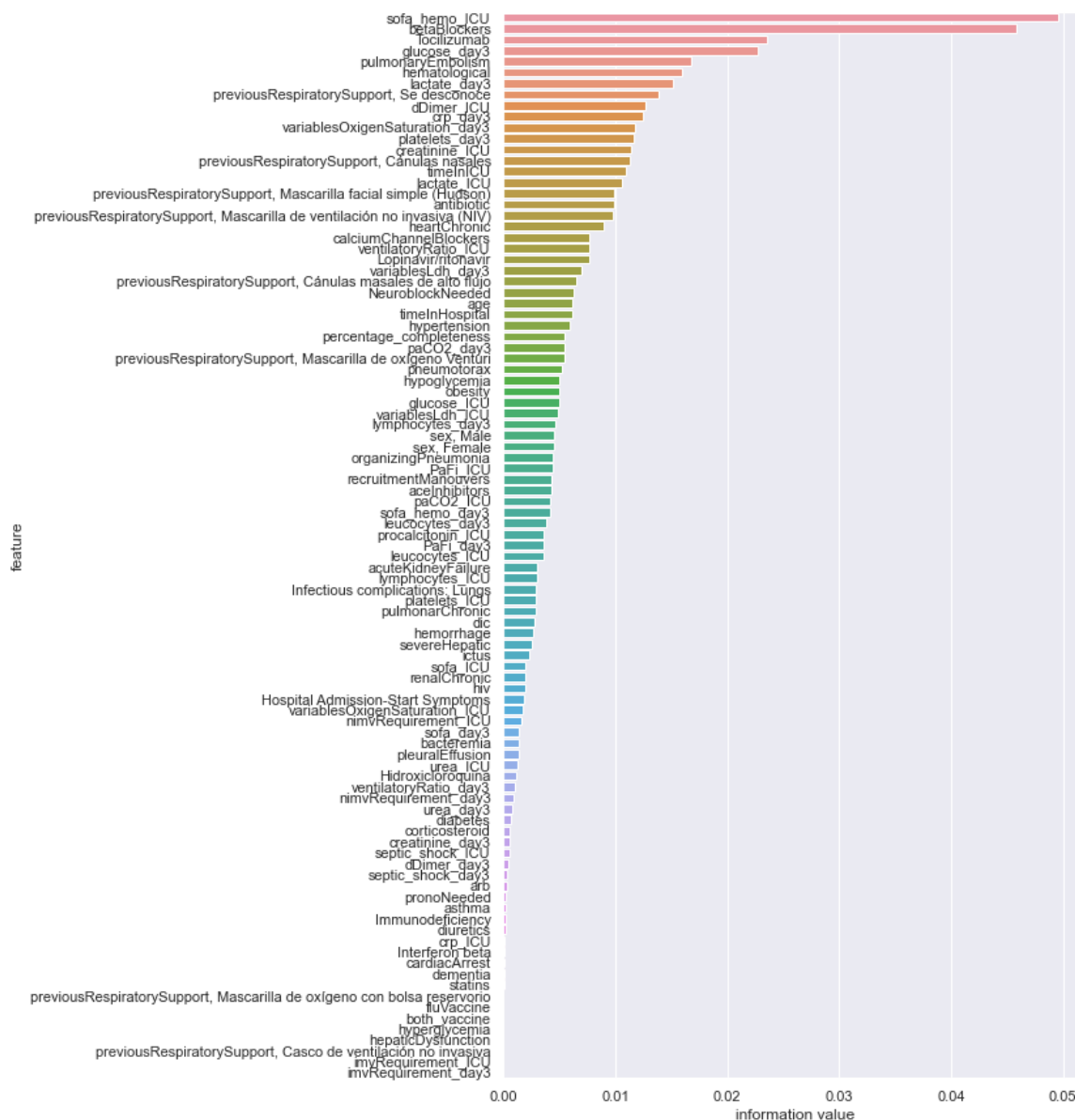


Figure 10.121: Information value feature selection with imputed features

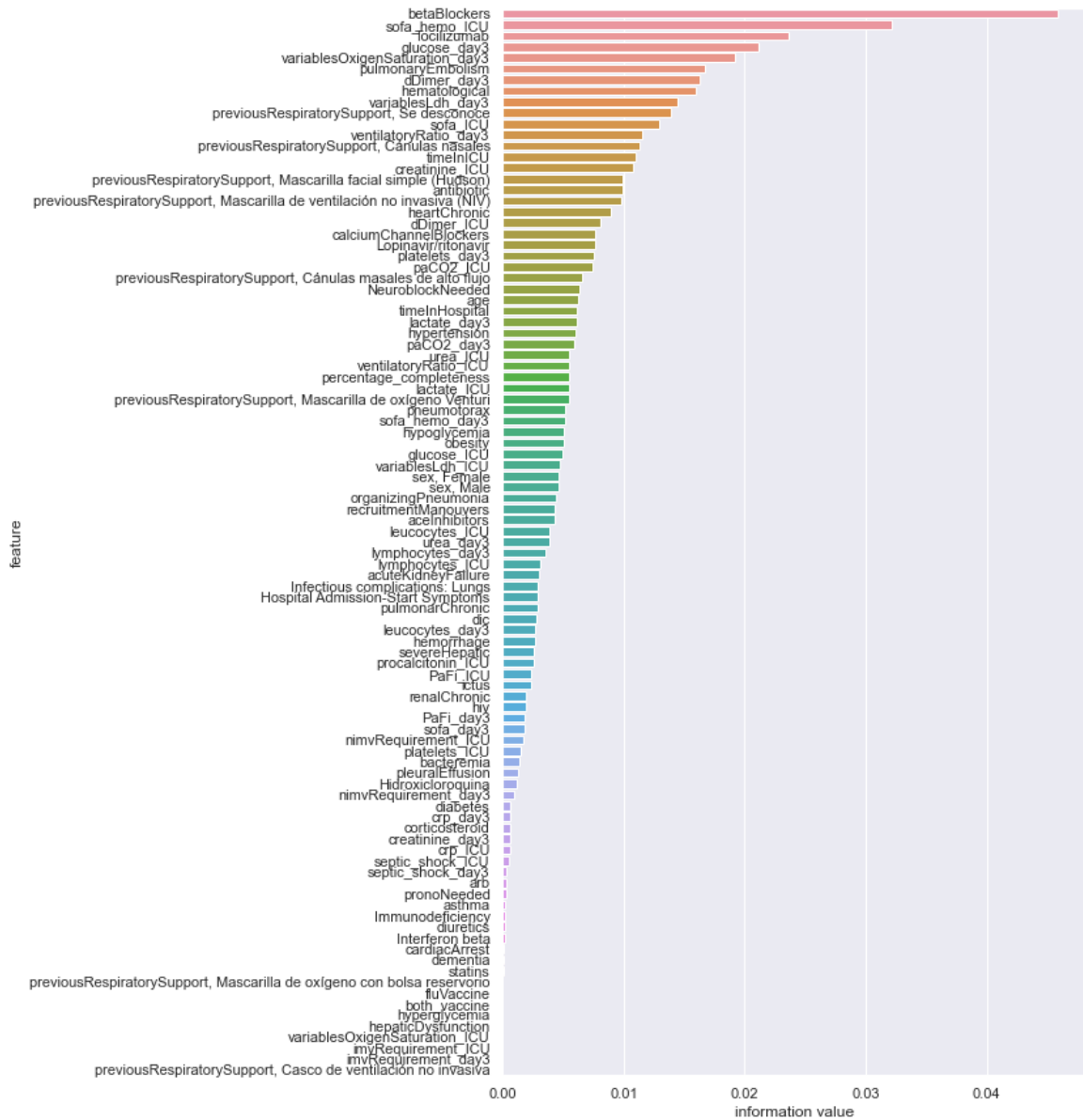


Figure 10.122: Information value feature selection

Weight?	Feature
+2.000	equal_ICU_IMV_times
+0.001	previousRespiratorySupport, Casco de ventilación no invasiva
+0.000	hematological
+0.000	Interferon beta
+0.000	hemorrhage
+0.000	statins
+0.000	Infectious complications: Lungs
	... 13 more positive ...
	... 12 more negative ...
-0.000	septic_shock_ICU
-0.000	obesity
-1.000	<BIAS>

Figure 10.123: Resulting weights for SVM linear kernel in approach A

Weight?	Feature
+38.000	equal_ICU_IMV_times
+8.000	renalChronic
+6.000	hemorrhage
+4.261	sofa_day3
+4.000	pneumotorax
	... 10 more positive ...
	... 11 more negative ...
-4.000	obesity
-4.057	creatinine_ICU
-5.085	variablesOxygenSaturation_ICU
-5.304	platelets_ICU
-23.000	<BIAS>

Figure 10.124: Resulting weights for perceptron in approach A

Weight?	Feature
+5.129	equal_ICU_IMV_times
+0.719	renalChronic
+0.527	statins
+0.522	pneumotorax
+0.424	bacteremia
+0.412	septic_shock_ICU
+0.398	hemorrhage
+0.309	creatinine_day3
	... 16 more positive ...
	... 9 more negative ...
-0.655	obesity
-2.970	<BIAS>

Figure 10.125: Resulting weights for logistic regression in approach A

column_1	column_2	correlation_coeff
timeInICU	timeInIMV	0.8482
timeInIMV	timeInICU	0.8482
creatinine_ICU	creatinine_day3	0.6097
creatinine_day3	creatinine_ICU	0.6097
ventilatoryRatio_ICU	ventilatoryRatio_day3	0.4678
ventilatoryRatio_day3	ventilatoryRatio_ICU	0.4678
sofa_day3	creatinine_day3	0.4513
creatinine_day3	sofa_day3	0.4513
lactate_ICU	septic_shock_ICU	0.4419
septic_shock_ICU	lactate_ICU	0.4419
renalChronic	creatinine_ICU	0.3872
creatinine_ICU	renalChronic	0.3872
crp_day3	crp_ICU	0.3708
crp_ICU	crp_day3	0.3708

Figure 10.126: Correlated variables in approach A [1/2]

column_1	column_2	correlation_coeff
timeInICU	timeInIMV	0.8482
timeInIMV	timeInICU	0.8482
creatinine_ICU	creatinine_day3	0.6097
creatinine_day3	creatinine_ICU	0.6097
ventilatoryRatio_ICU	ventilatoryRatio_day3	0.4678
ventilatoryRatio_day3	ventilatoryRatio_ICU	0.4678
sofa_day3	creatinine_day3	0.4513
creatinine_day3	sofa_day3	0.4513
lactate_ICU	septic_shock_ICU	0.4419
septic_shock_ICU	lactate_ICU	0.4419
renalChronic	creatinine_ICU	0.3872
creatinine_ICU	renalChronic	0.3872
crp_day3	crp_ICU	0.3708
crp_ICU	crp_day3	0.3708

Figure 10.127: Correlated variables in approach A [2/2]

Weight?	Feature
+2.023	equal_ICU_IMV_times
+0.464	bacteremia
+0.355	acuteKidneyFailure
+0.353	corticosteroid
+0.299	Infectious complications: Lungs
+0.193	Lopinavir/ritonavir
...	40 more positive ...
...	85 more negative ...
-0.198	day_159.0
-0.334	fluVaccine
-0.804	<BIAS>
-1.000	ICU_discharge

Figure 10.128: Resulting weights for SVM linear kernel in approach B

Weight?	Feature
+6013.000	age
+3157.214	timeInIMV
+2008.376	leucocytes_day3
+1985.000	equal_ICU_IMV_times
+1778.392	crp_day3
+1079.867	crp_ICU
... 18 more positive ...	
... 106 more negative ...	
-1409.273	Hospital Admission-Start Symptoms
-1494.154	PaFi_ICU
-2861.583	platelets_day3
-8085.114	timeInICU

Figure 10.129: Resulting weights for perceptron in approach B

Weight?	Feature
+0.270	equal_ICU_IMV_times
+0.249	timeInIMV
+0.072	acuteKidneyFailure
... 20 more positive ...	
... 105 more negative ...	
-0.067	day_16.0
-0.068	day_11.0
-0.069	day_15.0
-0.070	day_12.0
-0.070	day_14.0
-0.070	day_13.0
-0.237	timeInICU

Figure 10.130: Resulting weights for logistic regression in approach B

	column_1	column_2	correlation_coeff
	timeInIMV	timeInICU	0.8482
	timeInICU	timeInIMV	0.8482
	urea_day3	creatinine_day3	0.6680
	creatinine_day3	urea_day3	0.6680
	creatinine_day3	creatinine_ICU	0.6097
	creatinine_ICU	creatinine_day3	0.6097
	platelets_day3	platelets_ICU	0.5960
	platelets_ICU	platelets_day3	0.5960
	creatinine_ICU	urea_day3	0.4682
	urea_day3	creatinine_ICU	0.4682
	acuteKidneyFailure	creatinine_day3	0.4394
	creatinine_day3	acuteKidneyFailure	0.4394
	diabetes	glucose_ICU	0.4315
	glucose_ICU	diabetes	0.4315
	sofa_hemo_ICU	sofa_hemo_day3	0.3835

Figure 10.131: Correlated variables in approach B [1/3]

sofa_hemo_day3	sofa_hemo_ICU	0.3835
Interferon beta	Lopinavir/ritonavir	0.3738
Lopinavir/ritonavir	Interferon beta	0.3738
crp_ICU	crp_day3	0.3708
crp_day3	crp_ICU	0.3708
urea_day3	acuteKidneyFailure	0.3706
acuteKidneyFailure	urea_day3	0.3706
diabetes	glucose_day3	0.3691
glucose_day3	diabetes	0.3691
bacteremia	timeInIMV	0.3604
timeInIMV	bacteremia	0.3604
timeInICU	bacteremia	0.3514
bacteremia	timeInICU	0.3514
timeInICU	Infectious complications: Lungs	0.3370
Infectious complications: Lungs	timeInICU	0.3370
glucose_ICU	glucose_day3	0.3362

Figure 10.132: Correlated variables in approach B [2/3]

glucose_day3	glucose_ICU	0.3362
timeInIMV	Infectious complications: Lungs	0.3354
Infectious complications: Lungs	timeInIMV	0.3354
Infectious complications: Lungs	bacteremia	0.3246
bacteremia	Infectious complications: Lungs	0.3246
creatinine_ICU	acuteKidneyFailure	0.3152
acuteKidneyFailure	creatinine_ICU	0.3152
sofa_hemo_day3	septic_shock_day3	0.3083
septic_shock_day3	sofa_hemo_day3	0.3083

Figure 10.133: Correlated variables in approach B [3/3]

Bibliography

- [1] A. Torres, M. Arguimbau, J. Bermejo-Martín, R. Campo, A. Cecato, L. Fernandez-Barat, R. Ferrer, N. Jarillo, J. Á. Lorente-Balanza, R. Menéndez *et al.*, “Ciberesucicovid: A strategic project for a better understanding and clinical management of covid-19 in critical patients,” *Archivos de Bronconeumología*, 2020. 1
- [2] T. Chen, D. Wu, H. Chen, W. Yan, D. Yang, G. Chen, K. Ma, D. Xu, H. Yu, H. wu Wang, T. Wang, W. Guo, J. Chen, C. Ding, X. Zhang, J. Huang, M. Han, S. Li, X. Luo, J. Zhao, and Q. Ning, “Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study,” *The BMJ*, vol. 368, 2020. 3, 4
- [3] J. Casas-Rojo, J. M. Antón-Santos, J. Millán-Núñez-Cortés, C. Lumbreras-Bermejo, J. Ramos-Rincón, E. Roy-Vallejo, A. Artero-Mora, F. Arnalich-Fernández, J. M. García-Bruñén, J. Vargas-Núñez, S. Freire-Castro, L. Manzano-Espinosa, I. Perales-Fraile, A. Crestelo-Vieitez, F. Puchades-Gimeno, E. Rodilla-Sala, M. N. Solís-Marquínez, D. Bonet-Tur, M. Fidalgo-Moreno, E. Fonseca-Aizpuru, F. J. Carrasco-Sánchez, E. Rabadán-Pejenaute, M. Rubio-Rivas, J. D. Torres-Peña, and R. Gómez-Huelgas, “Clinical characteristics of patients hospitalized with covid-19 in spain: results from the semi-covid-19 registry,” *Revista Clinica Espanola*, vol. 220, pp. 480 – 494, 2020. 4, 30, 31
- [4] C. Kaeuffer, C. Hyaric, T. Fabacher, J. Mootien, B. Dervieux, Y. Ruch, A. Hugerot, Y.-J. Zhu, V. Pointurier, R. Clere, V. Greigert, L. Kassegne, N. Lefebvre, F. Gallais, A. Covid, S. Group, N. Meyer, Y. Hansmann, O. Hinschberger, and F. Danion, “Clinical characteristics and risk factors associated with severe covid-19: prospective analysis of 1,045 hospitalised cases in north-eastern france, march 2020,” *Eurosurveillance*, 09 2020. 4, 30, 31

- [5] M. Colaneri, P. Sacchi, V. Zuccaro, S. Biscarini, M. Sachs, S. Roda, T. Pieri, P. Valsecchi, A. Piralla, E. Seminari, A. Matteo, S. Novati, L. Maiocchi, L. Pagnucco, M. Tirani, F. Baldanti, F. Mojoli, S. Perlini, and R. Bruno, “Clinical characteristics of coronavirus disease (covid-19) early findings from a teaching hospital in pavia, north italy, 21 to 28 february 2020,” *Eurosurveillance*, vol. 25, 04 2020. 4, 30
- [6] J. Piñana, R. Martino, I. García-García, R. Parody, M. Morales, G. Benzo, I. Gómez-Catalan, R. Coll, I. Fuente, A. Luna, B. Merchán, A. Chinea, D. Miguel, A. Serrano, C. Pérez, C. Diaz, J. Lopez Lorenzo, A. Sáez, R. Bailen, and A. Sureda, “Risk factors and outcome of covid-19 in patients with hematological malignancies,” *Experimental Hematology Oncology*, vol. 9, 12 2020. 4
- [7] J.-U. A. H.-R. J. [U+FFFD] N. B. N. L. A. C. C. R. P. F. J. C. J. E. P. Q.-C. K. R. G. R. V. A. M. R. D.-G. A. B.-P. L. G.-S. E. L.-J. J. M. A. H. B.-O. M. Regalado-Artamendi, I., “Risk factors and mortality of covid-19 in patients with lymphoma: A multicenter study,” *HemaSphere*, vol. 25, 12 2021. 4
- [8] X. Li, S. Xu, M. Yu, K. Wang, Y. Tao, Y. Zhou, J. Shi, M. Zhou, B. Wu, Z. Yang, C. Zhang, J. Yue, Z. Zhang, H. Renz, X. Liu, J. Xie, M. Xie, and J. ping Zhao, “Risk factors for severity and mortality in adult covid-19 inpatients in wuhan,” *The Journal of Allergy and Clinical Immunology*, vol. 146, pp. 110 – 118, 2020. 4
- [9] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong, Y. Zhao, Y. Li, X. Wang, and Z. Peng, “Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in wuhan, china.” *JAMA*, 2020. 4
- [10] Z. Wu and J. McGoogan, “Characteristics of and important lessons from the coronavirus disease 2019 (covid-19) outbreak in china: Summary of a report of 72 314 cases from the chinese center for disease control and prevention,” *JAMA*, vol. 323, 02 2020. 4
- [11] W. Guan, Z. yi Ni, Y. Hu, W. Liang, C. Ou, J. He, L. Liu, H. Shan, C. Lei, D. Hui, B. Du, L. Li, G. Zeng, K.-Y. Yuen, R. Chen, C. Tang, T. Wang, P. Chen, J. Xiang, S. Li, J. lin Wang, Z. jing Liang, Y. xiang Peng, L. Wei, Y. Liu, Y. hua Hu, P. Peng, J. ming Wang, J. yang Liu, Z. Chen, G. Li, Z. jian Zheng, S. Qiu, J. Luo, C. Ye, S. yong Zhu, and

- N. Zhong, "Clinical characteristics of 2019 novel coronavirus infection in china," *medRxiv*, 2020. 4
- [12] R. Huang, L. Zhu, L. Xue, L. Liu, X. bing Yan, J. Wang, B. Zhang, T. min Xu, F. Ji, Y. Zhao, J. Cheng, Y. Wang, H. Shao, S. Hong, Q. Cao, C. Li, X. Zhao, L. Zou, D. Sang, H. Zhao, X. Guan, X. Chen, C. Shan, J. Xia, Y. Chen, X. Yan, J. Wei, C. Zhu, and C. Wu, "Clinical findings of patients with coronavirus disease 2019 in jiangsu province, china: A retrospective, multi-center study," *PLoS Neglected Tropical Diseases*, vol. 14, 2020. 4
- [13] Z. Zheng, F. Peng, B. Xu, J. Zhao, H. Liu, J. Peng, Q. Li, C. Jiang, Y. Zhou, S. Liu, C. Ye, P. Zhang, Y. Xing, H. Guo, and W. Tang, "Risk factors of critical mortal covid-19 cases: A systematic literature review and meta-analysis," *Journal of Infection*, vol. 81, 04 2020. 4
- [14] Z. B.-T. M. N.-Y. K. Z. Y. . Z. S. Chen, L., "Clinical course of severe and critically ill patients with coronavirus disease 2019 (covid-19): A comparative study," *The Journal of infection*, 2020. 4
- [15] W. Spain and M. Jané, "The first wave of the covid-19 pandemic in spain: Characterisation of cases and risk factors for severe outcomes, as at 27 april 2020," *Eurosurveillance*, vol. 25, 12 2020. 4
- [16] D. Bertsimas, G. Lukin, L. Mingardi, O. Nohadani, A. Orfanoudaki, B. Stellato, H. Wiberg, S. González-García, C. Parra-Calderón, K. Robinson, M. Schneider, B. Stein, A. Estirado, L. a Beccara, R. Canino, M. D. Bello, F. Pezzetti, and A. Pan, "Covid-19 mortality risk assessment: An international multi-center study," *PLoS ONE*, vol. 15, 2020. 4
- [17] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, M. Zhang, X. Huang, H. Cao, Y. Chen, T. Ren, F. Wang, Y. Xiao, S. Huang, X. Tan, N. Huang, and Y. Yuan, "An interpretable mortality prediction model for covid-19 patients," *Nature Machine Intelligence*, vol. 2, pp. 1–6, 05 2020. 4
- [18] P. A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J. G. Conde, "Research electronic data capture (redcap)—a metadata-driven methodology and workflow process for providing translational research informatics support," *Journal of biomedical informatics*, vol. 42, no. 2, pp. 377–381, 2009. 6

- [19] P. A. Harris, R. Taylor, B. L. Minor, V. Elliott, M. Fernandez, L. O’Neal, L. McLeod, G. Delacqua, F. Delacqua, J. Kirby *et al.*, “The redcap consortium: Building an international community of software platform partners,” *Journal of biomedical informatics*, vol. 95, p. 103208, 2019. 6
- [20] Michael Widenius, *MariaDB*, MariaDB Foundation, 2009. [Online]. Available: <https://mariadb.com/> 8
- [21] K. Pearson, “Notes on regression and inheritance in the case of two parents,” *Proceedings of the Royal Society of London*, 1895. 13
- [22] R. Fisher, “Statistical methods for research workers,” *Springer, New York*, 1992. 13
- [23] H. Cramer, “Mathematical methods of statistics,” *Princeton university Press*, 1946. 13
- [24] Shaked Zychlinski, *Dython*, BSD. [Online]. Available: <http://shakedzy.xyz/dython/> 14
- [25] H. Kang, “The prevention and handling of the missing data,” *Korean journal of anesthesiology*, vol. 64, pp. 402–6, 05 2013. 16
- [26] J. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, “When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts,” *BMC Medical Research Methodology*, vol. 17, 12 2017. 19
- [27] M. Azur, E. Stuart, C. Frangakis, and P. Leaf, “Multiple imputation by chained equations: What is it and how does it work?” *International journal of methods in psychiatric research*, vol. 20, pp. 40–9, 03 2011. 19, 20
- [28] A. Rubinsteyn and S. Feldman, *fancyimpute v0.5.5*, 2020. [Online]. Available: <https://pypi.org/project/fancyimpute/> 20
- [29] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, “API design for machine learning software: experiences from the scikit-learn project,” in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122. 23, 50

- [30] S. Fotso *et al.*, “PySurvival: Open source package for survival analysis modeling,” 2019–. [Online]. Available: <https://www.pysurvival.io/> 23, 50
- [31] A. Berkel and N. Siddiqi, “Building loss given default scorecard using weight of evidence bins in sas® enterprise miner™ loss given default,” 04 2012. 25
- [32] J.-F. Xie, W. Wu, S. Li, Y. Hu, M. Hu, J. Li, Y. Yang, T. Huang, K. Zheng, Y. Wang, H. Kang, Y. Huang, L. Jiang, W. Zhang, M. Zhong, L. Sang, X. Zheng, C. Pan, R. Zheng, and B. Du, “Clinical characteristics and outcomes of critically ill patients with novel coronavirus infectious disease (covid-19) in china: A retrospective multicenter descriptive study,” *SSRN Electronic Journal*, 01 2020. 30, 31
- [33] T. Clark, M. Bradburn, S. Love, and D. Altman, “Survival analysis part i: Basic concepts and first analyses,” *British journal of cancer*, vol. 89, pp. 232–8, 08 2003. 32, 33, 34
- [34] “Survival analysis – statistical methods and how we can use them for effective decision making,” 2020, coditation.com [Online]. 33
- [35] M. Goel, P. Khanna, and J. Kishore, “Understanding survival analysis: Kaplan-meier estimate,” *International journal of Ayurveda research*, vol. 1, pp. 274–8, 10 2010. 34, 35
- [36] D. Altman, “Practical statistics for medical research,” 1990. 34
- [37] V. Stel, F. Dekker, G. Tripepi, C. Zoccali, and K. Jager, “Survival analysis i: The kaplan-meier method,” *Nephron. Clinical practice*, vol. 119, pp. c83–8, 06 2011. 34
- [38] M. Bradburn, T. Clark, S. Love, and D. Altman, “Survival analysis part ii: Multivariate data analysis – an introduction to concepts and methods,” *British Journal of Cancer*, vol. 89, pp. 431 – 436, 2003. 36
- [39] S. Dessai and V. Patil, “Testing and interpreting assumptions of cox regression analysis,” *Cancer Research, Statistics, and Treatment*, vol. 2, p. 108, 01 2019. 36
- [40] X. Xue, X. Xie, M. Gunter, T. Rohan, S. Wassertheil-Smoller, G. Ho, D. Cirillo, H. Yu, and H. Strickler, “Testing the proportional hazards assumption in case-cohort analysis,” *BMC medical research methodology*, vol. 13, p. 88, 07 2013. 36

- [41] Z. Zhang, J. Reinikainen, K. Adeleke, M. Pieterse, and C. Groothuis-Oudshoorn, “Time-varying covariates and coefficients in cox regression models,” *Annals of Translational Medicine*, vol. 6, pp. 121–121, 04 2018. 37, 43
- [42] P. Pinsky, C. S. Zhu, and B. Kramer, “Lung cancer risk by years since quitting in 30+ pack year smokers,” *Journal of Medical Screening*, vol. 22, pp. 151 – 157, 2015. 37
- [43] J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*, 2nd ed. John Wiley & Sons, 2002. 37
- [44] D. Witten and R. Tibshirani, “Survival analysis with high-dimensional covariates,” *Statistical methods in medical research*, vol. 19, pp. 29–51, 09 2009. 38
- [45] J. Hao, Y. Kim, T. Mallavarapu, J. H. Oh, and M. Kang, “Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data,” *BMC Medical Genomics*, vol. 12, p. 189, 12 2019. 38
- [46] D. Faraggi and R. Simon, “A neural network model for survival data.” *Statistics in medicine*, vol. 14 1, pp. 73–82, 1995. 38
- [47] J. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, “Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network,” *BMC Medical Research Methodology*, vol. 18, 2018. 38
- [48] H. Ishwaran, U. Kogalur, E. Blackstone, and M. Lauer, “Random survival forests,” *The Annals of Applied Statistics*, vol. 2, 12 2008. 39
- [49] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, “Bias in random forest variable importance measures: Illustrations, sources and a solution.” *bmc bioinformatics*, 8(1), 25,” *BMC bioinformatics*, vol. 8, p. 25, 02 2007. 39
- [50] M. Wright, T. Dankowski, and A. Ziegler, “Unbiased split variable selection for random survival forests using maximally selected rank statistics,” *Statistics in Medicine*, vol. 36, 01 2017. 40
- [51] T. Hothorn, K. Hornik, and A. Zeileis, “Unbiased recursive partitioning: A conditional inference framework,” *Journal of Computational and Graphical Statistics*, vol. 15, pp. 651–674, 09 2006. 40

- [52] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Machine Learning*, vol. 63, pp. 3–42, 04 2006. 40
- [53] V. Van Belle and S. Huffel, “Support vector machines for survival analysis,” *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare*, 01 2007. 41
- [54] S. Pölsterl, N. Navab, and A. Katouzian, “Fast training of support vector machines for survival analysis,” 09 2015, pp. 243–259. 41
- [55] P. Austin, D. Lee, and J. Fine, “Introduction to the analysis of survival data in the presence of competing risks,” *Circulation*, vol. 133, pp. 601–609, 02 2016. 42, 43
- [56] Rebecca Scherzer, “A tutorial on accounting for competing risks in survival analysis,” 05 2017. 42
- [57] C. Davidson-Pilon, J. Kalderstam, N. Jacobson, S. Reed, B. Kuhn, P. Zivich, M. Williamson, AbdealJK, D. Datta, A. Fiore-Gartland, A. Parij, D. Wilson, Gabriel, L. Moneda, A. Moncada-Torres, K. Stark, H. Gadgil, Jona, JoseLlanes, K. Singaravelan, L. Besson, M. S. Peña, S. Anton, A. Klintberg, GrowthJeff, J. Noorbakhsh, M. Begun, R. Kumar, S. Hussey, and S. Seabold, “Camdavidsonpilon/lifelines: 0.26.0,” May 2021. [Online]. Available: <https://doi.org/10.5281/zenodo.4816284> 50
- [58] Mikhail Korobov, Konstantin Lopuhin , *ELI5*, MIT, 2016. 53
- [59] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: <https://www.R-project.org> 68
- [60] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate - a practical and powerful approach to multiple testing,” pp. 289 – 300, 11 1995.
- [61] C. An, H. Lim, D. wook Kim, J. Chang, Y. Choi, and S. W. Kim, “Machine learning prediction for mortality of patients diagnosed with covid-19: a nationwide korean cohort study,” *Scientific Reports*, vol. 10, 2020.