MSc Information Systems

Final Year Project

Predicting mortality in Intensive Care Units with artificial intelligence

By

Davide Garofalo

Project unit: PJS60

Supervisor: Jim Briggs

September 2020

Computing and Mathematics Programme Area

Table of Contents

Abstract 0
1 Introduction 3
1.1 Problem 3
1.2 Project's aim & objectives 3
1.3 Ethical issues involved 4
1.4 Structure of the dissertation 5
2 Project Management 7
2.1 Requirement analysis 7
2.2 Crisp-DM 7
2.3 Project planning 9
3 Research background 11
3.1 Introduction 11
3.2 Data Mining Techniques 11
3.2.1 Data Exploration 12
3.2.2 Data Preprocessing 12
3.2.3 Classification 12
3.3 Vital Signs monitoring and EWSs 13
3.4 Adverse Outcomes prediction with Machine Learning and Artificial Intelligence 13
4 The MIMIC-III database 17
4.1 Introduction 17
4.2 Tables and attributes description 18
4.3 A matter of time 21
5 Understanding the data 23
5.1 Data extraction 23
First steps 23
Getting patient related data 23
5.2 Cohort selection & construction 23
Calculating exclusion criteria flags 23
Getting outcomes columns 24
5.3 Cohort descriptive analysis 24
5.4 Exploratory Data Analysis 26

5.4.1 Cohort demographic distributions....27
5.4.2 Gender related analysis....29
5.4.3 Distribution of frequency of vital signs measurement during the day....30
5.4.4 Single feature distributions....31
5.4.5 Vital signs....32
5.4.6 Laboratory results....37
6 Data Pre-processing....45
6.1 The problem of time intervals....45
6.2 Dealing with outliers....53
6.3 Dealing with missing values....54
6.4 Other data transformations....55
7 Modelling....57
7.1 Feature sets....57
7.2 Balanced algorithms....58
7.2.1 Bagging....58
7.2.2 Boosting....59
7.2.3 Balanced Random Forest....59
7.3 Models evaluation....60
8 Conclusions....63
8.1 Recommendations for future improvements....63
References....65
Appendix A1: Project Initiation Document....71
Appendix A2: Gantt Diagram....77
Appendix B: Certificate of Ethics review....80
Appendix C: CITI certificate....82

List of Figures

Figure 2-1 - Crisp-DM phases 8
Figure 4-1 - MIMIC-III database structure 18
Figure 4-2 - Full list of MIMIC-III tables 19
Figure 5-1 - Number of observations excluded for the final cohort 24
Figure 5-2 - Distribution of age for the selected cohort 27
Figure 5-3 - Distribution of gender for the selected cohort (left) and confidence intervals for binomial proportion of gender (right) 28
Figure 5-4 - Distribution of the length of stay for the selected cohort 28
Figure 5-5 - Mortality distribution breaking down the population by age (x axis) and gender (Females on the left and Males on the right) 29
Figure 5-6 – Frequency of measurements of some vital signs during the day 30
Figure 5-7 –Heart Rate distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right) 32
Figure 5-8 – Mean Blood Pressure distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right) 33
Figure 5-9 – Body Temperature (*C) distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right) 34
Figure 5-10 – Respiratory Rate distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right) 35
Figure 5-11 – Oxygen Saturation distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right) 36
Figure 5-12 – Haemoglobin distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right) 37
Figure 5-13 – White Blood Cell count distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right) 38
Figure 5-14 – Blood Urea Nitrogen distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right) 39

Figure 5-15 – Albumin distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right).....40
Figure 5-16 – Creatinine distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right) 41
Figure 5-17 – Sodium distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right).....42
Figure 5-18 – Potassium distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right) 43
Figure 6-1 – SpO2 recording time intervals with summary statistics.....46
Figure 6-2 – Diastolic Blood Pressure recording time intervals with summary statistics 46
Figure 6-3 – Systolic Blood Pressure recording time intervals with summary statistics.....47
Figure 6-4 – Heart Rate recording time intervals with summary statistics.....47
Figure 6-5 – Body Temperature (°C) recording time intervals with summary statistics 48
Figure 6-6 – Glucose recording time intervals with summary statistics.....49
Figure 6-7 – Albumin recording time intervals with summary statistics 49
Figure 6-8 – Blood Urea Nitrogen recording time intervals with summary statistics.....50
Figure 6-9 – Creatinine recording time intervals with summary statistics.....50
Figure 6-10 – Hemoglobin recording time intervals with summary statistics 51
Figure 6-11 – White Blood Cell count recording time intervals with summary statistics 51
Figure 6-12 – Sodium recording time intervals with summary statistics.....52
Figure 6-13 – Potassium recording time intervals with summary statistics 52
Figure 7-1 – Final feature set shape.....57
Figure 7-2 – Example of randomized search (Balanced Random Forest).....58
Figure 7-3 – Confusion matrices of the best performing classifiers (RUSBoost on the left - Balanced Random Forest on the right).....61

List of Tables

Table 4-1 - Admissions table 19
Table 4-2 - Patients table 20
Table 4-3 - Chartevents table 20
Table 5-1 - Vital signs descriptive analysis: it includes summary statistics, unit measures, NA(missing) values count, values count and ranges (limits) of allowed values. 25
Table 5-2 - Laboratory results descriptive analysis: it includes summary statistics, unit measures, NA(missing) values count, values count and ranges (limits) of allowed values. 26
Table 6-1 - Plausible extreme value ranges 53
Table 6-2 - Number of outliers (Vital Signs) 54
Table 6-3 - Number of outliers' data (Laboratory Results) 54
Table 7-1 – Different configuration of the dataset by combination of features and missing value's filling strategy 57
Table 7-2 – Last pre-processing steps and total remaining # of rows 57
Table 7-3 – Bagging classifier parameters (optimized by randomized search) 59
Table 7-4 – Easy Ensemble classifier parameters (optimized by randomized search) 59
Table 7-5 – RUSBoost classifier parameters (optimized by randomized search) 59
Table 7-6 – BalancedRandomForest classifier parameters (optimized by randomized search) 60
Table 7-7 – Classification scores for selected algorithms 60

Abstract

This project aims to be a support for decisions in medical settings, particularly in Intensive Care Units. Predicting mortality by the use of statistics, Data Mining Techniques (DMT) and Artificial Intelligence (AI) can give to the medical practitioner deeper insight in the patient's actual condition, especially when time for action is limited: this can give them sufficient additional information to adequately intensify targeted care and optimize hospital resources.

With mortality being one of the most general outcomes to predict, this study was conducted on a general population of adult patients, examining their 24-hour vital sign data, laboratory test results, and demographics.

The main focus of the project has been on the data exploration, cleaning and handling techniques to reach a good accuracy for the prediction of ICU's patients' mortality at the 24h mark. The best results have been obtained by the application of balanced Machine Learning models. The highest balanced accuracy in prediction reached is 71.5%.

Word Count: 11731

1 Introduction

1.1 Problem

Predicting mortality in ICU is something that, if improved, not only can save more lives by escalating targeted care for who needs it but can also save big sums of money for public healthcare by optimizing hospital resources and nurse workload.

Despite being one of the most common clinical practices, vital sign monitoring is still a field open for improvement on multiple sides. In recent times, the adoption of early warning scores has been highly effective for detecting patients at risk of clinical deterioration or death, prompting a timelier clinical response, with the aim of improving patient outcomes in the NHS.

In the early eighties, it was estimated that an average ICU patient is described by about 250 different parameters (Goldman, 2015). A typical human capacity can handle not more than 5-7 different parameters at a time (Goldman, 2015; Ramon et al., 2007). As a result, a data-driven system can help assist clinicians to detect problems earlier than an experienced intensivist would (Silva et al., 2006). The challenge then is to learn which patterns indicate which problems and learn to predict these patterns in the huge amount of data available as early as possible.

Therefore, this study aims to investigate the use of Data Mining Techniques (DMT) and Artificial Intelligence (AI) classification models in the realm of early mortality prediction.

The research is conducted on an anonymous secondary dataset obtained through the MIMIC-III database (*MIMIC Critical Care Database*, n.d.): database comprises detailed clinical information regarding more than 60 000 stays in Intensive Care Units (ICUs) at the Beth Israel Deaconess Medical Centre in Boston, Massachusetts, collected as part of routine clinical care between 2001 and 2012.

1.2 Project's aim & objectives

The aim of this project is to build an artificial intelligence model that has good accuracy in predicting ICU mortality. The objectives that must be achieved to reach this final goal are:

- Explore the MIMIC-III database and getting useful insights
- Through intense literature research, acquire basic knowledge of vital signs monitoring and existing artificial intelligence techniques used in this field
- Extract a meaningful cohort that could be representative for the general population
- Apply DMT to pre-process the data and enhance its quality
- Design, build and validate machine learning and deep learning models to predict mortality in ICU
- Evaluate the final results and discuss possible future improvements

1.3 Ethical issues involved

Generally, risks associated with record-based research (as this project) stem from possible invasion of privacy and breaches of confidentiality.

In the context of research, privacy risk mainly relates to the techniques used to collect information about subjects. Privacy risks are relatively small when subjects actually consent to provide personal information. However, privacy risks are much higher when researchers obtain information for research purposes without the subject's consent, which is often the case in records-based research.

Confidentiality relates to the actual handling of the personal information once it is obtained and how will it be used, stored, and reported in a way that is consistent with the manner under which it was originally obtained from the individual.

The risks of breach of confidentiality associated with records-based research are generally linked to the sensitivity of the requested information. The security of the information is less of a concern when the information is registered without identifiers. If the information is both identifiable and sensitive, methods to protect confidentiality must be carefully considered.

If we analyse those two risks in the context of the MIMIC database, we can easily deduce that they're actually low, but still, careful security measures should be applied.

Before data was incorporated into the MIMIC-III database, it was first deidentified in accordance with Health Insurance Portability and Accountability Act (HIPAA) standards using structured data cleansing and date shifting. The de identification process for structured data required the removal of all eighteen of the identifying data elements listed in HIPAA, including fields such as patient name, telephone number, address, and dates. In particular, dates were shifted into the future by a random offset for each individual patient in a consistent manner to preserve intervals. Time of day, day of the week, and approximate seasonality were conserved during date shifting.

Protected health information was removed from free text fields, such as diagnostic reports and physician notes, using a rigorously evaluated de identification system based on extensive dictionary lookups and pattern-matching with regular expressions. The components of this deidentification system are continually expanded as new data is acquired.

The MIMIC project was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Centre (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA).

Moreover, to use the data for research purposes, one should complete a training course about USA ethics and legislation for human research. After the training, you can request the access by signing an informed consent in which you agree to:

- Not attempt to identify any individual or institution referenced in the restricted data.
- Exercise all reasonable and prudent care to avoid disclosure of identities and maintain the physical and electronic security of the data.
- Not share access to the data with anyone else.
- Use the data for research purposes only.

1.4 Structure of the dissertation

This dissertation is composed of 8 chapters and its structure is organized as follows:

- Chapter 2 – Project Management: This chapter describes the six phases of CRISP-DM (Cross-Industry Process for Data Mining) process with which the project is being planned to execute.
- Chapter 3 – Research background: This chapter introduces the relevant background to this dissertation, both Data Mining and clinical related. Most of this chapter focus on reviewing clinical literature about Early Warning Score systems and adverse outcome prediction with Machine Learning and Deep Learning techniques
- Chapter 4 – The MIMIC-III database: This chapter describes in depth the database used for the study
- Chapter 5 – Understanding the data: This chapter presents all the different stages of statistical analysis and visualization used to better understand the data. The steps taken to extract the study cohort are described and all relevant distributions are plotted for each of the dataset's feature
- Chapter 6 – Data Pre-processing: This chapter describes the relevant pre-processing techniques applied to reach the best quality possible for the extracted data. Strategies for dealing with missing values and outliers are discussed in depth
- Chapter 7 – Modelling: This chapter describes the balanced algorithms used for the mortality prediction and evaluates their performances
- Chapter 8 – Conclusions: This chapter describes the outcomes of the research. Limitations and future work are also presented

2 Project Management

2.1 Requirement analysis

In order to start this project, many different types of requirements have been analyzed and broken down: functional requirements, requirements of skills, knowledge and tools.

Starting from the knowledge, it has been acquired through intensive literature search (see chapter 3) and includes two main themes:

- Healthcare related knowledge, regarding vital sign monitoring, its values and practices, the usage of aggregated early warning scores etc.
- Artificial intelligence and data mining related knowledge, regarding the best models used in recent times for similar classification problems and the best practices used to handle health related data.

The tools used for the development of the project included my personal laptop (for prototyping, writing reports and deliverables, knowledge research etc.), a cloud-based querying framework with an instance of the MIMIC-III database for the most expensive queries (Google Big Query), a cloud-based computing platform (Google Colab) for training the AI models, softwares for big data analysis and data mining (R Studio, Python, Jupyter Notebook and PostgreSQL).

Main skills essential for the success of the project, either already possessed or acquired during the development phase are:

1. SQL querying
2. Python data pre-processing (using libraries such as NumPy, pandas, matplotlib)
3. R for data visualization and exploration
4. Basic statistical analysis
5. Features selection and extraction techniques
6. Resampling and other class balancing techniques
7. Artificial intelligence algorithms for binary and probabilistic classification

Functional requirement that dictated the success of the project has been a reasonably good prediction rate of the AI models (65% - 70% accuracy or more and an Area Under the Receiver Operator Curve (AUROC) of 70% or more).

2.2 Crisp-DM

To optimize the most important and scarce resource of the project (time) I followed the CRISP-DM methodology (Wirth & Hipp, 2000) and a strict and accurate planning schedule. CRISP-DM stands for Cross Industry Standard Process for Data Mining and is a 1996 methodology created to shape Data Mining projects. It consists of 6 steps to conceive a Data Mining project and they can have cycle iterations according to developers' needs. It is still widely used today (*CRISP-DM, Still the Top Methodology for Analytics, Data Mining, or Data Science Projects*, n.d.) because it is an open standard that is easy to follow and highly effective.

1. Set goals for the Big Data project	1. Business Understanding
2. Set the data and data sources	2. Data Understanding
3. Check if the available data can meet the objectives of the project and establish how you will meet the objectives	2. Data Understanding
4. The Data is transformed, in the cases that is necessary for the Big Data process	3. Data Preparation
5. Execute the algorithms that satisfies the project objectives	4. Modeling
6. The results are presented, analyzed and disseminated.	5. Evaluation and 6 .Deployment
7. Depending on the results, strategic decisions are taken to follow.	5. Evaluation and 6 .Deployment

Figure 2-1 - Crisp-DM phases

The six phases of CRISP-DM:

1. Business Understanding: first phase of the CRISP-DM process which focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.
2. Data Understanding: second phase of the CRISP-DM process which focuses on data collection, checking quality and exploring data to discover first insights into the data to form hypotheses for hidden information.
3. Data Preparation: third phase of the CRISP-DM process which focuses on selection and preparation of final dataset. This phase may include many tasks, such as table, records and attributes selection as well as cleaning and transformation of data.
4. Modelling: fourth phase of the CRISP-DM process which focuses on selection and application of various modelling techniques. Different parameters are set, and different models are built for the same data mining problem.
5. Evaluation: fifth phase of the CRISP-DM process which focuses on the evaluation of obtained models and deciding on how to use the results. Interpretation of the model depends upon the algorithm. Models can be evaluated to review whether they achieve the objectives or not.
6. Deployment: final phase of the CRISP-DM process which focuses on determining the use of the obtained knowledge and results. This phase also focuses on organizing, reporting and presenting the gained knowledge when needed.

2.3 Project planning

Main strategy for project planning has been writing down a work breakdown structure and estimate the time needed for every activity: the final product of this first planning phase is the original Gantt chart attached in the appendix A1.

Note that the end of the project has in reality been delayed by two weeks due to the difficulties encountered during the Covid-19 pandemic.

The fundamental resource for rescheduling and replanning activities has been a weekly meeting with supervisors.

3 Research background

3.1 Introduction

The primary concern of any healthcare system is to relieve the patient's symptoms, prevent complications and prolong the patient's life. In order to achieve these goals, it is crucial in the ICU to provide the correct treatment and to predict clinical deterioration early enough so preventive or curative actions can be taken in time (Awad et al., 2017).

The most widely used ICU outcome measure is mortality, as it is patient-centered, objective and easily measured up to the point of hospital discharge (Higgins, 2007): if timely and accurately predicted, could save lives and money. To achieve this prevention goal, many different approaches have been proposed and used in recent years, the most used being the routine monitoring of vital signs and laboratory results in order to obtain a more complete knowledge of the current patient's state of health.

According to Ramon et al. (Ramon et al., 2007), 70% of ICU patients need vital support only for a few days and have high chance of survival, while 30% of patients stay in the ICU for a longer period, sometimes three weeks or more, which is associated with greater mortality.

The use of ICU data in the early prediction of mortality is an attractive area for investigation, both for reasons of quality and cost; it has the potential to be of value in the assessment of severity of illness and adjustment of healthcare policy.

Identifying patients at risk and patients who might not benefit from ICU treatment is crucial as the ICU is very costly with limited resources. In addition, the ICU is a data-rich environment where critically ill patients are constantly monitored with complex equipment inputting hundreds of medical measurements every day. As result, this makes it a well-suited setting for the implementation of an automated data-driven system which analyzes large amounts of raw data, that could be overlooked by human experts, and extracts high-level information in order to predict in-hospital mortality early on.

In the world of big data technologies, the general framework that is used for pattern recognition is the one of Data Mining Techniques.

3.2 Data Mining Techniques

Data Mining (DM) is an analytic process designed to explore data in large data repositories in order to find novel and useful patterns that might otherwise remain hidden. The existing patterns, associations and relationships among data can be exploited and transformed into actionable knowledge. Technically, Data Mining is the process of finding correlations or patterns in datasets where the overall goal is to extract information from a dataset and transform it into an understandable format for further use (Lewicki & Hill, 2006).

For this reason, Data Mining Techniques are the core of most AI application: data is everything and everything is data, to allow a model to learn something useful from a sea of raw data we need

to carefully remove what is unnecessary and to clean and aggregate what is really useful and high in entropy.

3.2.1 Data Exploration

Managing big data, with a large number of dimensions and a huge variety of data types, can be quite challenging - that's why the first step in a good approach to data mining is knowing your data. Data exploration is primarily useful for familiarizing yourself with the dataset and helping you investigate the behavior of the data: it is useful for understanding the complexity of the data, identifying inaccuracies, analyzing the presence of outliers, data types and scales of data.

The two main tools for data exploration are descriptive statistics and data visualization, both of which complement each other. With the visualization of the data it is possible to see the distribution of each feature or explore high-dimensional data, obtaining the first insights and understanding. This will lead to the development of questions about the dataset that can be answered with additional visualization methods combined with descriptive statistics - such as five summary statistics (boxplots), statistical tests, etc. This process goes on and on as the analyst further explores the dataset until questions about the nature of the data are answered and / or more research questions are defined.

3.2.2 Data Preprocessing

After having thoroughly explored the dataset, the next step of Data Mining is Data Preprocessing. The purpose of this phase is to ensure that the data that will be used by the AI models is of the highest possible quality, because even the best and most complex model cannot learn meaningful patterns from chaotic and dirty data. Therefore, data preprocessing tasks range from mitigating noise, to balance unbalanced classes, fill missing values and incomplete data, reduce dimensionality and much more. Traditionally, it is divided in four main parts: data integration, cleaning, transformation and reduction.

Those topics will be discussed in depth in the preprocessing chapter.

3.2.3 Classification

Classification is one of the most researched problems in machine learning. It is a supervised machine learning approach that aims to identify an unseen object as part of a certain category or class. Classifiers learn from the dataset and build a model that can be used to predict a class based on the attributes alone. In this sense, the aim is to generate knowledge-based models which will help in predicting the behavior of new data. Classification methods are various and based on the application goal the method is chosen (Palmer et al., 2011).

Major classification techniques include Artificial Neural Networks (ANN) (Rumelhart et al., 1986), Bayesian classification (Giudici & Figini, 2009), Decision Trees (DTs) (Larose & Larose, 2014), Support Vector Machines (SVM) (Cortes & Vapnik, 1995), Naive Bayes (NB) (Han et al., 2012), Decision Rules (Montgomery, 2013) and K-Nearest Neighbor (KNN) (Peterson, 2009). These techniques make it possible to analyze categorical output variables in order to generate classification models.

3.3 Vital Signs monitoring and EWSs

In most hospital settings, vital signs include heart rate, temperature, respiratory rate and blood pressure. It has been proposed, however, that these four measurements may be combined with other indicators such as nutritional status, smoking status, level of consciousness, spirometry and pulse oximetry. Such variables are the easiest , cheapest and perhaps most important information that can be collected on hospitalized patients (Kellett & Sebat, 2017).

Even though vital signs were introduced into clinical practice more than a century ago, surprisingly few attempts were made to quantify their clinical performance (Kellett, 2017). Vital signs have been an area of intensive study in recent decades and several studies have found that alterations in vital signs occur several hours before a significant adverse event (Kause et al., 2004).

During recent years, UK researchers added inestimable value to this practice by the introduction of one kind of "track and trigger system" called aggregated Early Warning Score system (EWS) (Morgan et al., 1997). Six main vital signs (respiratory rate, oxygen saturation, temperature, diastolic and systolic blood pressure, pulse/heart rate, level of consciousness) are recorded for this method, then a score is calculated for each individual vital sign based on a predefined range and finally aggregated into a single EWS. Based on the calculated EWS, clinical care can be escalated appropriately (i.e. by calling a Rapid Response Team (RTT), increasing the frequency of measurements etc.) to increase the odds of a positive outcome.

Along with EWSs, a number of Severity of Illness scores have been introduced in the ICU to assess severity of illness. These include APACHE (Knaus et al., 1985), SAPS (Le Gall, 1993), MPM (Lemeshow et al., 1993), and SOFA score (Vincent et al., 1996). These scoring systems perform well at assessing patient risk with Areas Under the Receiver Operator Characteristics curve (AUROCs) typically between 0.8 and 0.9.

Research showed that the adoption of EWSs was highly effective for detecting patients at risk of clinical deterioration or death, prompting a timelier clinical response, with the aim of improving patient outcomes (Smith et al., 2013).

The success of EWSs is not only due to their better performances in predicting adverse outcomes (Jarvis et al., 2015; Prytherch et al., 2010) but is also addressable to the standardization process that gave birth to a National EWS (NEWS) in the United Kingdom. This standardization led to higher compliance to protocols and more lives saved.

A unified systematic and evidence-based approach has three major benefits:

- Achievement of faster and better results in medical wards.
- Staff should only need to be trained once and will adhere more to the protocols.
- Objective and standardized data of illness severity and clinical outcomes that can be exploited by future research.

3.4 Adverse Outcomes prediction with Machine Learning and Artificial Intelligence

Several authors have discussed ICU data including Ramon et al. (Ramon et al., 2007) who precisely described the various categories of ICU data. The first category of ICU data includes

parameter measurements, such as temperature, blood pressure, pulse…etc. Also, laboratory data from examination of samples or bacteriological data related to infections of patients and demographics are considered parameter measurements. The second category concerns treatment information; it includes medications administered to the patient, information about treatments the patient receives, feeding, and treatment policy decisions. The third category of is background data, which includes data on known effects of medications and the typical situations and reasons why they are applied. Expert knowledge is also considered background data.

In this study, data from the first category will be used because is the most common and readily available.

Various authors have advocated the use of Data Mining and Artificial Intelligence techniques for predicting ICU mortality over the use of simple logistic regression and SOI scores.

Research conducted by Dybowski et al. (Dybowski et al., 1996) and Henriques et al. (Henriques & Rocha, 2009) have reported better performance of Artificial Neural Network over logistic regression in predicting in-hospital mortality for critically ill patients. However, research conducted by Clermont et al., Wong et al. and Doig et al. (Clermont et al., 2001; Doig et al., 1993; Wong & Young, 1999) found that logistic regression and neural networks performed similarly for ICU mortality prediction. Others (Citi & Barbieri, 2012; Delen et al., 2005; Kim et al., 2011; Ribas et al., 2011) found that DTs and SVMs performed better. In 2011, Ribas et al. (Ribas et al., 2011) showed that the use of SVMs resulted in increased prediction accuracy as compared to the APACHE II score.

In addition, the work proposed by Sadeghi et al. in (Sadeghi et al., 2018) demonstrated the capability of the Random Forests classifier in terms of accuracy and interpretability in comparison with other methods in predicting mortality within the first hour of ICU admission using 12 features extracted from the heart signals of patients; similar results have been obtained by Churpek et al. (Churpek, Yuen, et al., 2016) while predicting various outcome with tree based algorithms. On the other hand, Ramon et al. (Ramon et al., 2007) reported that the AUROCs of DT based algorithms (DT learning, 65%; first order RF, 81%) yielded smaller areas compared to those of NB networks (AUROC, 85%) and tree-augmented NB networks (AUROC, 82%) in their study on a small dataset containing 1,548 mechanically ventilated ICU patients. Also, Pirracchio et al. (Pirracchio et al., 2015) reported that Bayesian Additive Regression Trees (BART) is the best candidate when using transformed variables, while Random Forests outperformed all other candidates when using untransformed variables.

For what concerns more complex models, Machine Learning algorithms produced superior classification performance in detection of sepsis onset compared with the most common scores (qSOFA, MEWS, SIRS, SAPS etc.) in the studies of Desautels et al. and Nemati et al. (Desautels et al., 2016; Nemati et al., 2018). Hofer et al. (Hofer et al., 2020), in a study on a population of perioperative patients, showed how Deep Neural Networks can be used to predict multiple outcomes using a single input feature set with better performance than the ASA physical status classification system.

The best performances in clinical outcomes prediction have been achieved with the implementation of Deep Learning models like Long Short term Memories (LStM), Recurrent

Neural Networks (RNN) etc. (Aczon et al., 2017; Harutyunyan et al., 2019; Jo et al., 2017). Although the existing studies utilizing RNN achieved higher accuracy of ICU mortality prediction compared with the traditional scoring systems based on logistic regression, they cannot provide explicit interpretability as scoring system can, and therefore lack face validity. Ge et al. and Barbieri et al. (Barbieri et al., 2020; Ge et al., 2018) made a step toward better interpretability of the results provided by those complex models.

From the studies mentioned above, it is clear that there is no single algorithm that outperforms others; it depends on the population of interest, the variables measured, and the outcome being tested. However, some models reveal strengths over others in certain aspects. For example, the major advantage for the use of DTs over other models lies in their descriptive modeling as they explain hidden clinical implications unlike Artificial Neural Networks (ANNs), which lacks logic between input and output nodes. From another perspective, DT, RF, ANN, BNs and kernel methods such as SVM can handle large size data samples and integrate background knowledge into analysis (Meyfroidt et al., 2009).

4 The MIMIC-III database

4.1 Introduction

The MIMIC-III database comprises detailed clinical information regarding more than 60 000 stays in ICUs at the Beth Israel Deaconess Medical Centre in Boston, Massachusetts, collected as part of routine clinical care between 2001 and 2012. The MIMIC-III dataset is freely available to researchers around the world and has been widely used in the development of predictive models, epidemiological studies, and educational courses. The database includes information such as demographics, vital sign measurements made at the bedside (~1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (including post-hospital discharge). Moreover, the MIMIC-III code repository (Johnson et al., 2018) offers a collection of freely-available resources and is intended to be a central hub for sharing, refining, and reusing code used for analysis of the MIMIC critical care database.

Reproducibility is a key factor in scientific studies and lack of it is a barrier to the generation of a robust knowledge base to support clinical decision-making. That's the main reason why a massive freely available data and code repository is so important. In addition, concepts are being defined collaboratively with those who are familiar with the workflows and clinical environment, including how the data are captured.

Even if the data has been collected in the USA, it is still highly relatable to other hospital settings, since vital sign monitoring is a common worldwide practice. Two different critical care information systems were in place over the data collection period: Philips CareVue Clinical Information System (models M2331A and M1215A; Philips Health-care, Andover, MA) and iMDsoft MetaVision ICU (iMDsoft, Needham, MA).

Notable features of the dataset that lead to its adoption for this project are:

- High temporal resolution: it is a key factor for simulating different frequency of observation while analyzing the data
- Large and diverse data population that includes not only vital signs but also things like Electronic Healthcare Records (EHR), laboratory test results and so on
- Highly active community of researcher sharing code and insights
- High reliability of the data due to the high-grade information systems used for collection
- Useful aggregated scores and concepts defined and computed by experts in the health-care sector

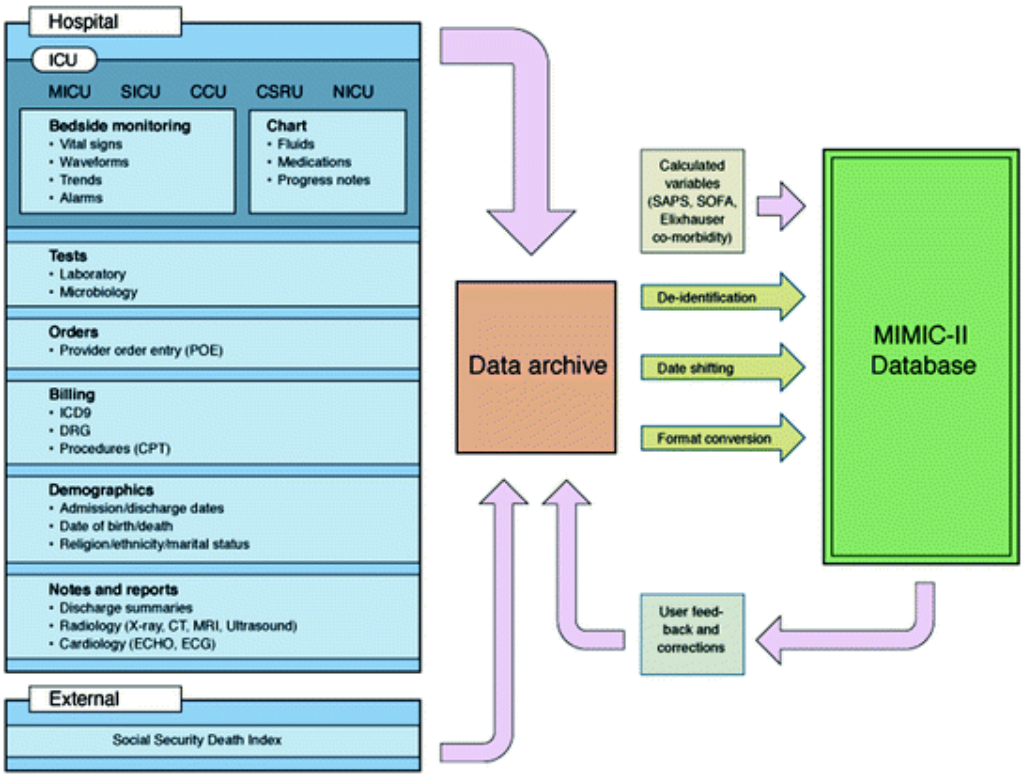


Figure 4-1 - MIMIC-III database structure

4.2 Tables and attributes description

MIMIC has 40 tables (Figure 4-1) that consist of hundreds of columns and millions of rows. A really good database design job has been made to reach the clear schema that it has today. Thanks to the good organization of the data within those tables, anyone who aims to use MIMIC for clinical research has to do queries most of the time only on the three main tables: PATIENTS, ADMISSIONS, CHARTEVENTS (as described below).

Table	Children	Parents	Columns	Rows	Comments
admissions	18	1	19	58,976	Hospital admissions associated with an ICU stay.
callout		2	24	34,499	Record of when patients were ready for discharge (called out), and the actual time of their discharge (or more generally, their outcome).
caregivers	7		4	7,567	List of caregivers associated with an ICU stay.
chartevents		5	15	330,712,483	Events occuring on a patient chart.
chartevents_1			15	38,033,561	Partition of chartevents. Should not be directly queried.
chartevents_10			15	9,584,888	Partition of chartevents. Should not be directly queried.
chartevents_11			15	470,141	Partition of chartevents. Should not be directly queried.
chartevents_12			15	265,413	Partition of chartevents. Should not be directly queried.
chartevents_13			15	39,066,570	Partition of chartevents. Should not be directly queried.
chartevents_14			15	100,075,138	Partition of chartevents. Should not be directly queried.
chartevents_2			15	13,116,197	Partition of chartevents. Should not be directly queried.
chartevents_3			15	38,657,533	Partition of chartevents. Should not be directly queried.
chartevents_4			15	9,374,587	Partition of chartevents. Should not be directly queried.
chartevents_5			15	18,201,026	Partition of chartevents. Should not be directly queried.
chartevents_6			15	28,014,688	Partition of chartevents. Should not be directly queried.
chartevents_7			15	255,967	Partition of chartevents. Should not be directly queried.
chartevents_8			15	34,322,082	Partition of chartevents. Should not be directly queried.
chartevents_9			15	1,274,692	Partition of chartevents. Should not be directly queried.
cptevents		2	12	573,146	Events recorded in Current Procedural Terminology.
d_cpt			9	134	High-level dictionary of the Current Procedural Terminology.
d_icd_diagnoses	1		4	14,710	Dictionary of the International Classification of Diseases, 9th Revision (Diagnoses).
d_icd_procedures	1		4	3,898	Dictionary of the International Classification of Diseases, 9th Revision (Procedures).
d_items	8		10	12,487	Dictionary of non-laboratory-related charted items.
d_labitems	1		6	753	Dictionary of laboratory-related items.
datetimeevents		5	14	4,485,937	Events relating to a datetime.
diagnoses_icd		3	5	651,047	Diagnoses relating to a hospital admission coded using the ICD9 system.
drgcodes		2	8	125,557	Hospital stays classified using the Diagnosis-Related Group system.
icustays	8	2	12	61,532	List of ICU admissions.
inputevents_cv		4	22	17,527,935	Events relating to fluid input for patients whose data was originally stored in the CareVue database.
inputevents_mv		5	31	3,618,991	Events relating to fluid input for patients whose data was originally stored in the MetaVision database.
labevents		3	9	27,854,055	Events relating to laboratory tests.
microbiologyevents		5	16	631,726	Events relating to microbiology tests.
noteevents		3	11	2,083,180	Notes associated with hospital stays.
outputevents		5	13	4,349,218	Outputs recorded during the ICU stay.
patients	19		8	46,520	Patients associated with an admission to the ICU.
prescriptions		3	19	4,156,450	Medicines prescribed.
procedureevents_mv		5	25	258,066	Procedure start and stop times recorded for MetaVision patients.
procedures_icd		3	5	240,095	Procedures relating to a hospital admission coded using the ICD9 system.
services		2	6	73,343	Hospital services that patients were under during their hospital stay.
transfers		3	13	261,897	Location of patients during their hospital stay.
40 Tables			**534**	**728,556,685**	

Figure 4-2 - Full list of MIMIC-III tables

The **ADMISSIONS table** gives information regarding a patient's admission to the hospital. Since each unique hospital visit for a patient is assigned a unique hospital admission ID number (HADM_ID), the ADMISSIONS table can be considered as a definition table for HADM_ID. Information available includes timing information for admission and discharge, demographic information, the source of the admission, and so on.

Name	Data type
ROW_ID	INT
SUBJECT_ID	INT
HADM_ID	INT
ADMITTIME	TIMESTAMP(0)
DISCHTIME	TIMESTAMP(0)
DEATHTIME	TIMESTAMP(0)
ADMISSION_TYPE	VARCHAR(50)
ADMISSION_LOCATION	VARCHAR(50)
DISCHARGE_LOCATION	VARCHAR(50)
INSURANCE	VARCHAR(255)
LANGUAGE	VARCHAR(10)
RELIGION	VARCHAR(50)
MARITAL_STATUS	VARCHAR(50)
ETHNICITY	VARCHAR(200)
DIAGNOSIS	VARCHAR(300)
HOSPITAL_EXPIRE_FLAG	TINYINT

Table 4-1 - Admissions table

The **PATIENTS table** defines every patient and assigns him or her a unique SUBJECT_ID. Here is defined patient's demographic information for the first time and these are then linked to the admission table. Date of birth (DOB) and date of death (DOD) are shifted to make it impossible tracing back patient's identity.

Name	Data type
ROW_ID	INT
SUBJECT_ID	INT
GENDER	VARCHAR(5)
DOB	TIMESTAMP(0)
DOD	TIMESTAMP(0)
DOD_HOSP	TIMESTAMP(0)
DOD_SSN	TIMESTAMP(0)
EXPIRE_FLAG	VARCHAR(5)

Table 4-2 - Patients table

The **CHARTEVENTS table** contains all the charted data available for a patient. During their ICU stay, the primary repository of a patient's information is their electronic chart. The electronic chart displays patients' routine vital signs and any additional information relevant to their care: ventilator settings, laboratory values, code status, mental status, and so on. As a result, the bulk of information about a patient's stay is contained in CHARTEVENTS. Furthermore, even though laboratory values are captured elsewhere (LABEVENTS), they are frequently repeated within CHARTEVENTS. This occurs because it is desirable to display the laboratory values on the patient's electronic chart, and so the values are copied from the database storing laboratory values to the database storing the CHARTEVENTS.

Name	Data type
ROW_ID	INT
SUBJECT_ID	NUMBER(7,0)
HADM_ID	NUMBER(7,0)
ICUSTAY_ID	NUMBER(7,0)
ITEMID	NUMBER(7,0)
CHARTTIME	DATE
STORETIME	DATE
CGID	NUMBER(7,0)
VALUE	VARCHAR2(200 BYTE)
VALUENUM	NUMBER
VALUEUOM	VARCHAR2(20 BYTE)

Table 4-3 - Chartevents table

4.3 A matter of time

In order to conform with the HIPAA standard and to maintain the anonymity and privacy of the patients, who collect the data took particular care to shift dates in a way that would preserve data consistency and meaning while rendering impossible to trace back patient's personal data. All dates in the database have been shifted to protect patient confidentiality. Dates will be internally consistent for the same patient, but randomly distributed in the future. This means that if measurement A is made at 2150-01-01 14:00:00, and measurement B is made at 2150-01-01 15:00:00, then measurement B was made 1 hour after measurement A.

The date shifting preserved the following:

- Time of day - a measurement made at 15:00:00 was actually made at 15:00:00 local standard time.
- Day of the week - a measurement made on a Sunday will appear on a Sunday in the future.
- Seasonality - a measurement made during the winter months will appear during a winter month.

The date shifting removed the following:

- Year - The year is randomly distributed between 2100 - 2200.
- Day of the month - The absolute day of the month is not preserved.
- Inter-patient information - Two patients in the ICU on 2150-01-01 were not in the ICU at the same time.

Dates of birth which occur in the present time are not true dates of birth. Furthermore, dates of birth which occur before the year 1900 occur if the patient is older than 89. In these cases, the patient's age at their first admission has been fixed to 300.

5 Understanding the data

5.1 Data extraction

Understanding the data that you are dealing with is the first and most important thing to do in every big data application: it allows the analyst to undertake the most appropriate actions in the pre-processing phase in order to reach the best possible data quality. This chapter will detail all the relevant steps taken from the collection of the data to its exploration.

First steps

To get access to the MIMIC-III download, any researcher needs to complete the MIT ethics course "Data or Specimens Only Research". The course comprised subjects like History and Ethics of Human Subjects Research, Research and HIPAA Privacy Protections, Conflicts of Interest in Human Subjects Research and so on. The completion of the course provided a certificate by CITI (Collaborative Institutional Training Initiative).

After downloading it, the MIMIC-III database has been installed on a local PostgreSQL server. The connection with the database has been established via R and Python's APIs. This installation was used to explore the data, get familiar with it and to do query prototyping. An important step to build domain specific knowledge has been studying the pre-built queries in the git-hub MIMIC-III code repository. Modifications of some of those queries have also been useful in the next step for extracting the selected study cohort.

Getting patient related data

Since data has been de-identified and time-shifted, patient's age is not explicit in the database and for this reason age has been extracted by subtracting their date of birth from the hospital admission date. Patient aged over 89 years had their age concealed by recording it with very large numbers: since the real median age of this group is known, their age has been substituted with the median value (91.4).

Other patient related data comprise IDs (admission ID, ICU stay ID, subject ID), gender, intime and outtime for ICU stays, admission and discharge time from the hospital.

5.2 Cohort selection & construction

Since the aim of the study is ICU mortality prediction in the most general terms possible, the study cohort has been constructed by applying only the mandatory exclusion criteria for this kind of clinical research.

Calculating exclusion criteria flags

Hospital Length of Stay (LOS) could be obtained by two means: extracted by subtracting the admission date from the discharge date or, getting it by the appropriate column in the Admissions table. The latter approach proved to be more suitable because the first could lead to inconsistencies due to possible known typographical errors in the date's recordings. A similar approach has been followed for obtaining the ICU length of stay.

The list of the inclusion criteria for the study cohort is the following:

- Age at admission >= 16 – The physiology of adults is substantially different from the one of children and adolescents
- Type of service not NB or NBB – Those two types of services refers to newborns
- First ICU stay for every hospital admission – Multiple ICU stay from the same admission tend to be highly correlated
- LOS >= 24h – Shorter stays could lead to insufficient data
- Admission type = “Urgent” or “Emergency” – Non elective admissions are less likely to end up in ICU

Getting outcomes columns

Last step of cohort definition is the extraction of the outcome variables. The main outcome variable is death in ICU. Both hospital and ICU LOS have been added to the dataset because they’re also very often used as outcome modelling variables and, during the exploration phase, have been useful to enrich the knowledge about the cohort.

```
Observations          61532
exclusion_age          8109 (13.18%)
exclusion_los         12308 (20.00%)
exclusion_first_stay   3746 (6.09%)
exclusion_serv         8088 (13.14%)
exclusion_admit_type  15364 (24.97%)

Total excluded        25551 (41.52%)

Final cohort observations: 35981
```

Figure 5-1 - Number of observations excluded for the final cohort

5.3 Cohort descriptive analysis

For the selected cohort’s ICU stays vital signs and most common laboratory analysis have been extracted using the cloud instance of MIMIC-III database on Google BigQuery (those queries were too expensive to be run locally). This instance of the database had concepts (already pre-aggregated data from different tables) implemented by a researcher with domain knowledge. Vital signs included Temperature (°C), SpO2, Diastolic BP, Systolic BP, Mean BP, Heart Rate, Respiratory Rate, GCS, Glucose.

Labs results included BUN, Haemoglobin, WBC, Creatinine, Albumin, Sodium, Potassium.

VITAL SIGN	SUMMARY	UNIT MEASURE	COUNT	LIMITS
Heart Rate	Min. : 0.3 1st Qu. : 74 Median : 86 Mean : 87.1 3rd Qu. : 99 Max. : 285	bpm Beats per Minute	Tot. : 4807082 (83.79%) NA's : 929893 (16.21%)	]0,300[
Respiratory Rate	Min. : 0.2 1st Qu. : 16 Median : 20 Mean : 20.3 3rd Qu. : 24 Max. : 69	Bpm Breaths per Minute	Tot. : 4831816 (84.22%) NA's : 905159 (15.78%)	]0,70[
Blood Pressure (Systolic)	Min. : 0.1 1st Qu. : 105 Median : 120 Mean : 122.2 3rd Qu. : 138 Max. : 341	mmHG	Tot. : 4439668 (77.38%) NA's : 1297307 (22.62%)	]0,400[
Blood Pressure (Diastolic)	Min. : 0.3 1st Qu. : 51 Median : 59 Mean : 60.7 3rd Qu. : 69 Max. : 298	mmHG	Tot. : 5607608 (97.74%) NA's : 129867 (2.26%)	]0,300[
Blood Pressure (Mean)	Min. : 0.4 1st Qu. : 68 Median : 78 Mean : 79 3rd Qu. : 89 Max. : 299	mmHG	Tot. : 4449522 (77.55%) NA's : 1287453 (22.45%)	]0,300[
SpO2	Min. : 0.5 1st Qu. : 96 Median : 98 Mean : 97 3rd Qu. : 99 Max. : 100	%	Tot. : 4689085 (81.73%) NA's : 1047890 (18.27%)	]0,100]
Temperature	Min. : 15 1st Qu. : 36 Median : 37 Mean : 37 3rd Qu. : 38 Max. : 43	°C	Tot. : 1315988 (22.93%) NA's : 4420987 (77.7%)	]10,50[
Glucose	Min. : 0 1st Qu. : 108 Median : 130 Mean : 145 3rd Qu. : 162 Max. : 999999	mg/dL	Tot. : 961472 (16.75%) NA's : 4775503 (83.25%)	]0, +inf[

Table 5-1 - Vital signs descriptive analysis: it includes summary statistics, unit measures, NA(missing) values count, values count and ranges (limits) of allowed values.

Vital signs data has 5736975 rows. We can spot some outliers by simply looking at the summaries. The *limits* that values can assume have been taken from the MIMIC-III code repository's concepts. Since there's no upper bound for Glucose values, some errors (marked as value 999999) have been carried over and will be removed when dealing with outliers. Also some extremely low values shown in the Min. of each vital sign could be possible outliers.

LABORATORY RESULT	SUMMARY	UNIT MEASURE	COUNT	LIMITS
Albumin	Min. : 0.9 1st Qu. : 2.4 Median : 2.9 Mean : 2.9 3rd Qu. : 3.4 Max. : 6.9	g/dL	Tot. : 81896 (9.48%) NA's : 781848 (90.52%)	]0,10[
Creatinine	Min. : 0.1 1st Qu. : 0.7 Median : 1 Mean : 1.6 3rd Qu. : 1.7 Max. : 138	mg/dL	Tot. : 494869 (57.29%) NA's : 368875 (42.71%)	]0,150[
Hemoglobin	Min. : 1.6 1st Qu. : 9.0 Median : 10.1 Mean : 10.3 3rd Qu. : 11.3 Max. : 22.3	g/dL	Tot. : 450428 (52.14%) NA's : 413316 (47.86%)	]0,50[
White Blood Cell count	Min. : 0.1 1st Qu. : 7 Median : 9.8 Mean : 11.3 3rd Qu. : 13.7 Max. : 846.7	Tens of thousands	Tot. : 450932 (52.20%) NA's : 412812 (47.80%)	]0,1000[
Serum Urea (BUN)	Min. : 1 1st Qu. : 14 Median : 23 Mean : 31 3rd Qu. : 40 Max. : 290	mg/dL	Tot. : 493151 (57.09%) NA's : 370593 (42.91%)	]0,300[
Potassium	Min. : 0.8 1st Qu. : 3.7 Median : 4 Mean : 4.1 3rd Qu. : 4.4 Max. : 27.5	mEq/L	Tot. : 525207 (60.80%) NA's : 338537 (39.20%)	]0,30]
Sodium	Min. : 74 1st Qu. : 136 Median : 139 Mean : 138.7 3rd Qu. : 142 Max. : 184	mEq/L	Tot. : 504983 (58.46%) NA's : 358761 (41.54%)	]0,200[

Table 5-2 - Laboratory results descriptive analysis: it includes summary statistics, unit measures, NA(missing) values count, values count and ranges (limits) of allowed values.

Laboratory results data has 863744 rows. The *limits* that values can assume have been taken from the MIMIC-III code repository's concepts. Some suspected outlier can be identified from the summaries (i.e. the Max value in WBC). Albumin has a low number of recordings (its implications will be discussed later).

5.4 Exploratory Data Analysis

In this section, we will visually and statistically explore all the main characteristics of the selected cohort of patients. In the first part we will look at numbers concerning general info about

demographics, whereas in the second part we'll look at specific distributions of the vital signs and laboratory analysis of the selected cohort.

5.4.1 Cohort demographic distributions

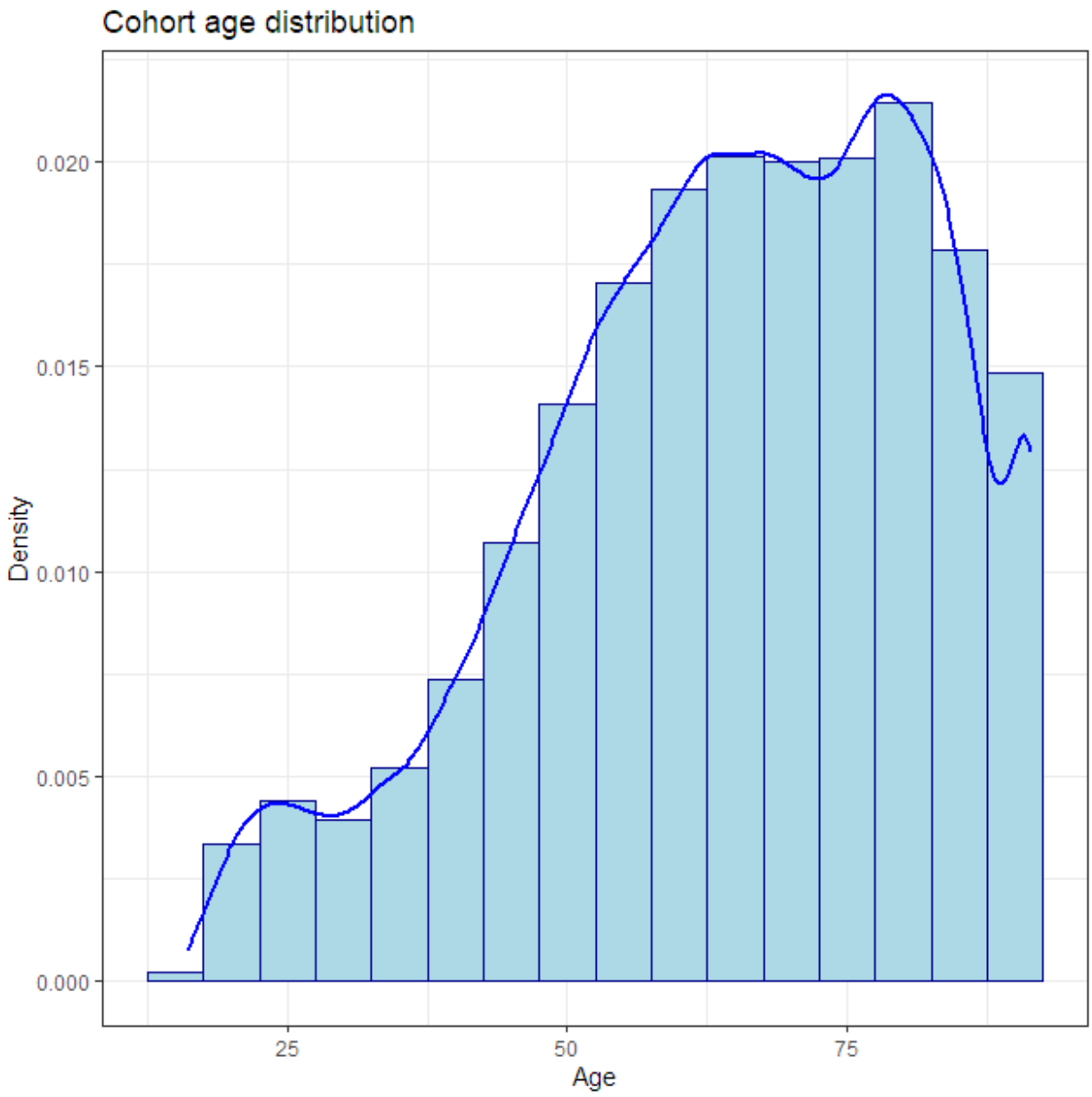
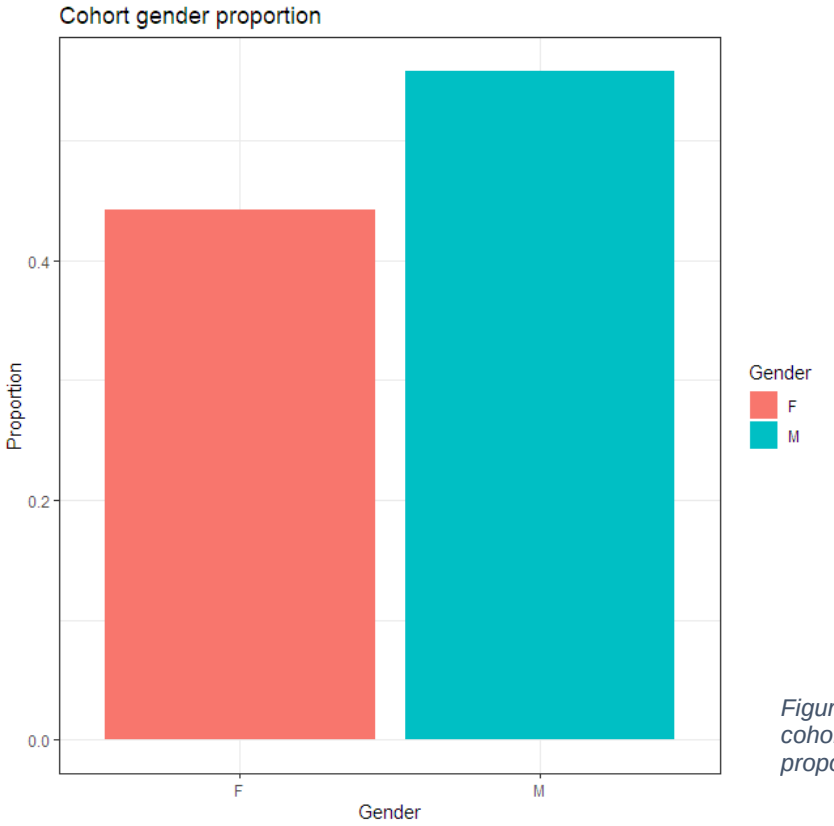


Figure 5-2 - Distribution of age for the selected cohort

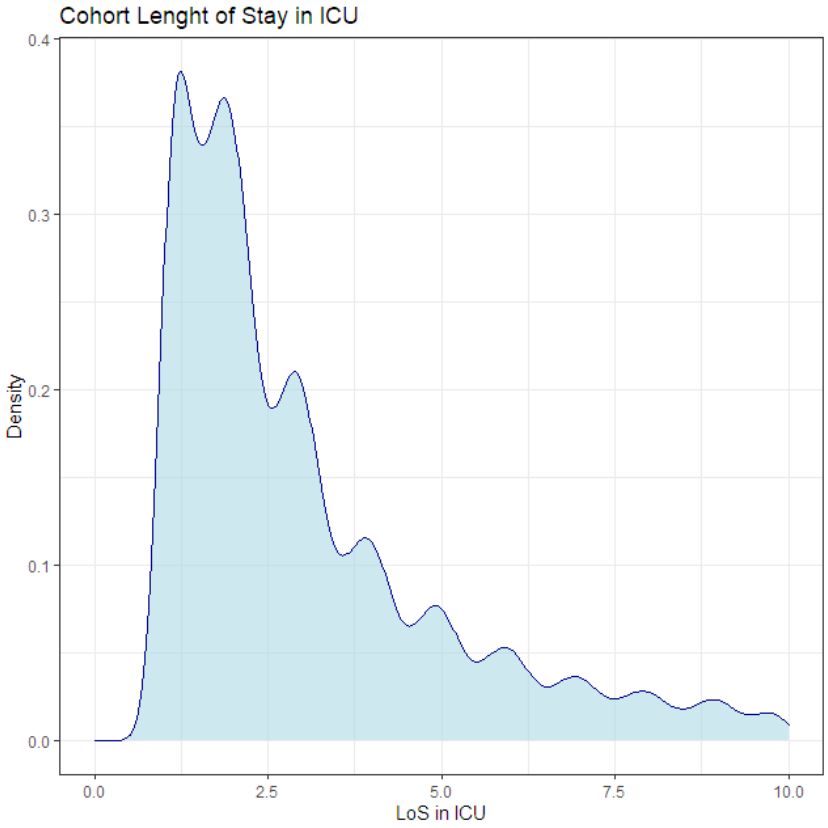
Patients have been filtered to be adults (>= 16 years of age). A large number of patients is aged 60+



Gender proportion for the cohort. Males are about 8% more numerous than women.

	est	lwr.ci	upr.ci
F	0.442	0.437	0.447
M	0.558	0.553	0.563

Figure 5-3 - Distribution of gender for the selected cohort (left) and confidence intervals for binomial proportion of gender (right)



Length of stay in ICU is skewed to the right, with a median of 2.68 days and a mean of 4.81 days. Minimum stay is fixed at 1 day by cohort exclusion criteria.

Figure 5-4 - Distribution of the length of stay for the selected cohort

5.4.2 Gender related analysis

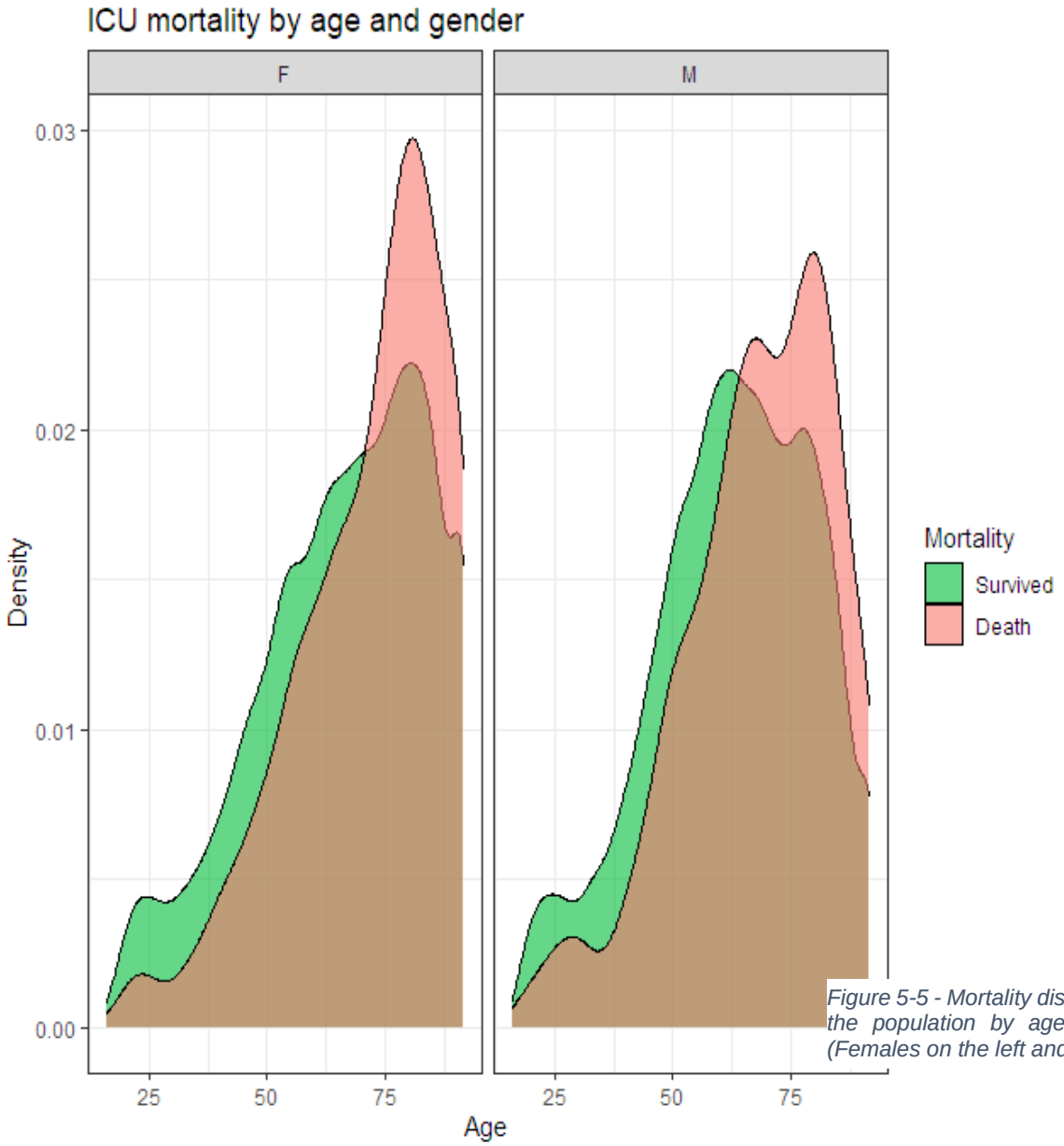


Figure 5-5 - Mortality distribution breaking down the population by age (x axis) and gender (Females on the left and Males on the right)

Females Mortality Proportion

	est	lwr.ci	upr.ci
survived_perc	0.909	0.905	0.914
death_perc	0.091	0.086	0.095

Females Mean Age of Death (with CI)

upper	mean	lower
72.13882	71.26090	70.38299

Males Mortality Proportion

	est	lwr.ci	upr.ci
survived_perc	0.919	0.915	0.924
death_perc	0.081	0.076	0.085

Males Mean Age of Death (with CI)

upper	mean	lower
68.42759	67.55958	66.69157

We can see that females tend to die slightly more in ICU (1% higher mortality) but have a mean life expectancy of almost 4 years higher than males. We can see from the t-test results that this difference is highly significant (p-value < 0.001).

```
> # Mean ICU LoS by gender
> # Mean ICU LoS for Females
> CI(onlyFemales$icu_los, ci = 0.95)
   upper     mean    lower
4.856681 4.759298 4.661915
> # Mean ICU LoS for Males
> CI(onlyMales$icu_los, ci = 0.95)
   upper     mean    lower
4.917562 4.815920 4.714277
```

No statistically significant difference is found in mean length of stay in ICU between males and females. (p-value: 0.43)

5.4.3 Distribution of frequency of vital signs measurement during the day

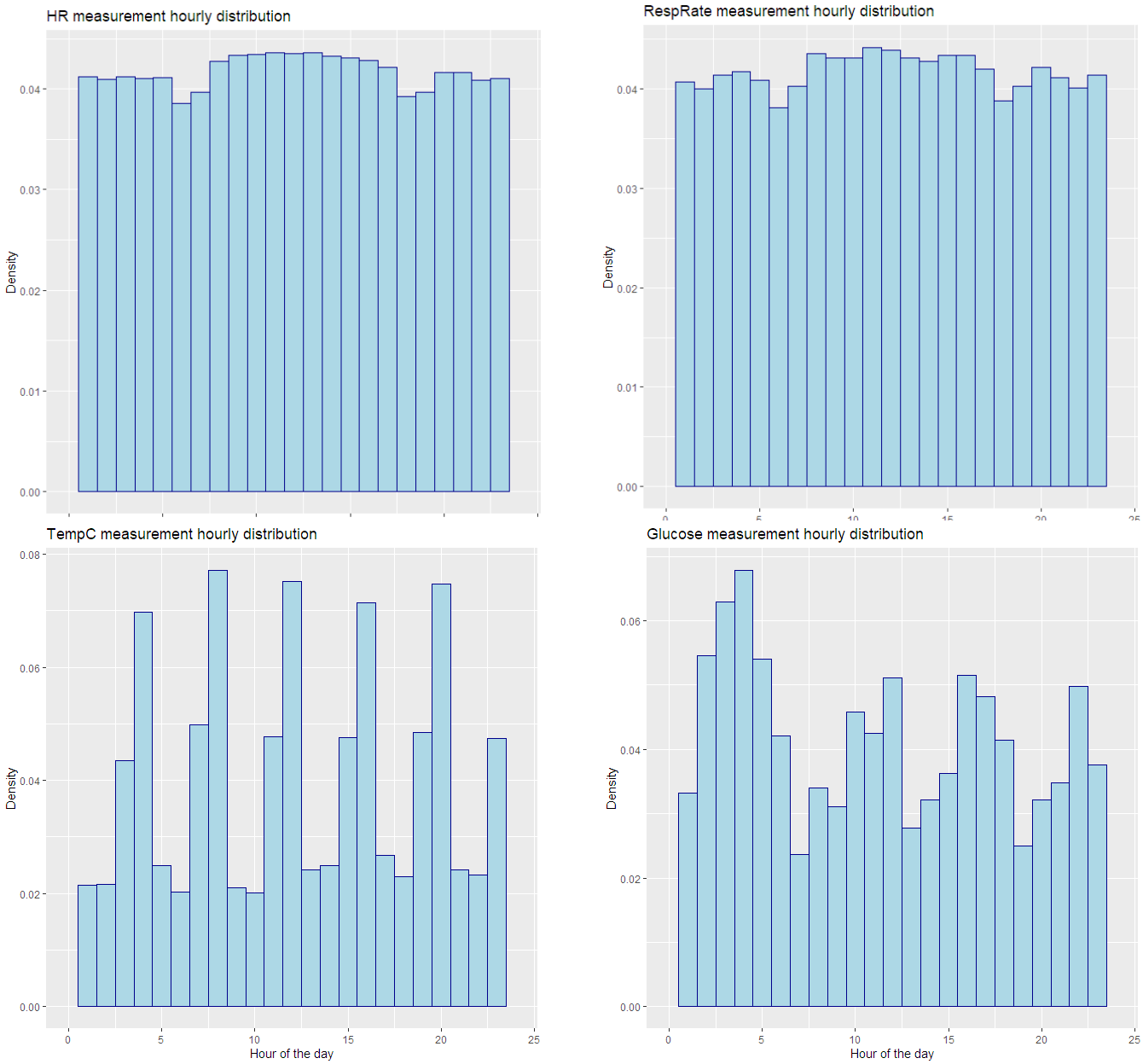


Figure 5-6 – Frequency of measurements of some vital signs during the day

As we can see from the previous graphs, HR and Respiratory Rate (BP and SpO2 have a similar pattern) are measured almost evenly during the day: slightly fewer measurements are present at times corresponding to the change of the nurses' shift.

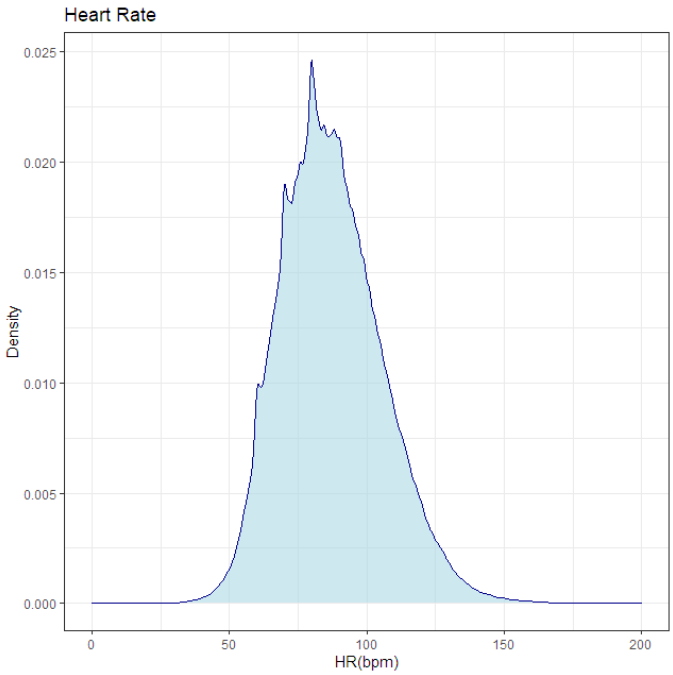
Temperature measurements are taken typically at 4-hour intervals. A similar pattern arises for glucose, but with higher concentration of measurements during the night.

5.4.4 Single feature distributions

In the following pages we'll look at the distributions of each feature (VS and Laboratory result) in the dataset. The layout of the pages will be the following: in the top left corner we have the distribution of the specific feature in the dataset, in the top right corner a description of that feature and on the bottom a ridge graph. The ridge graph shows mortality (colour coded) broken down for age group (Y axis) and gender (Males in the left pane and Females in the right one).

5.4.5 Vital signs

Heart Rate



Heart Rate or pulse rate is a measurement of the number of beats of the heart per minute. The normal pulse for healthy adults ranges from 60 to 100 beats per minute.
Elevated resting heart rate (RHR) is an independent risk factor for mortality (Jensen et al., 2013) as shown also from the graph below on the cohort data.

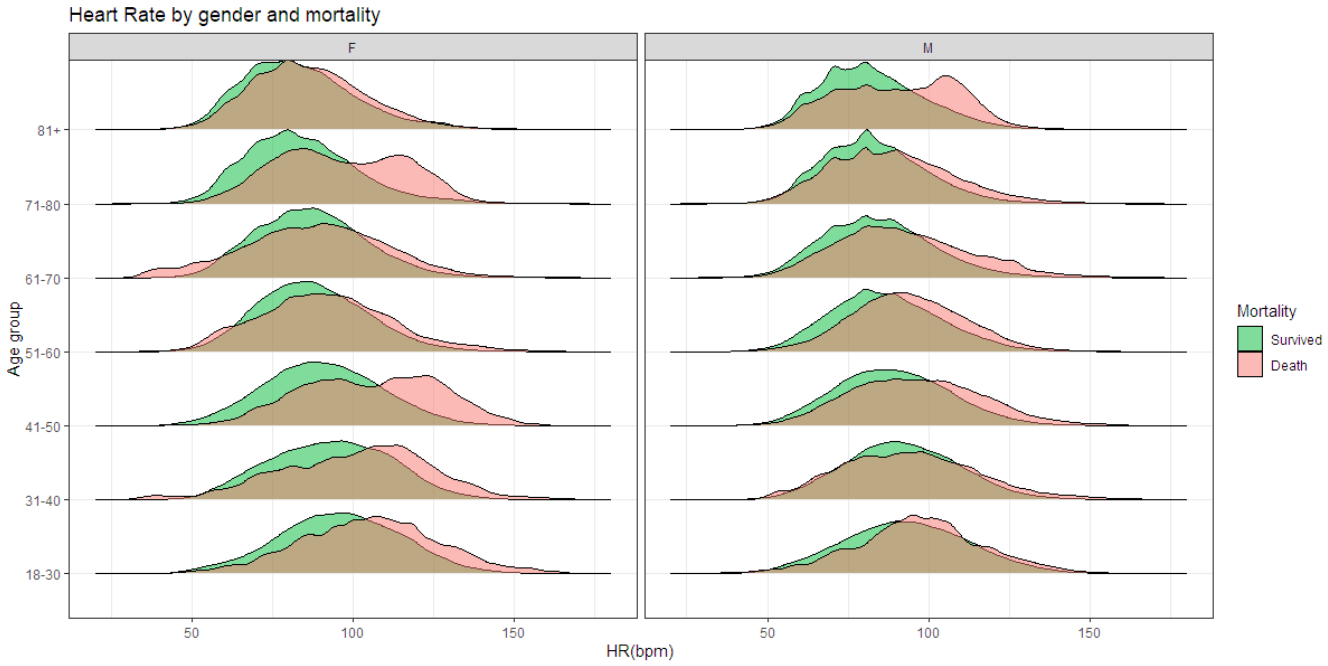
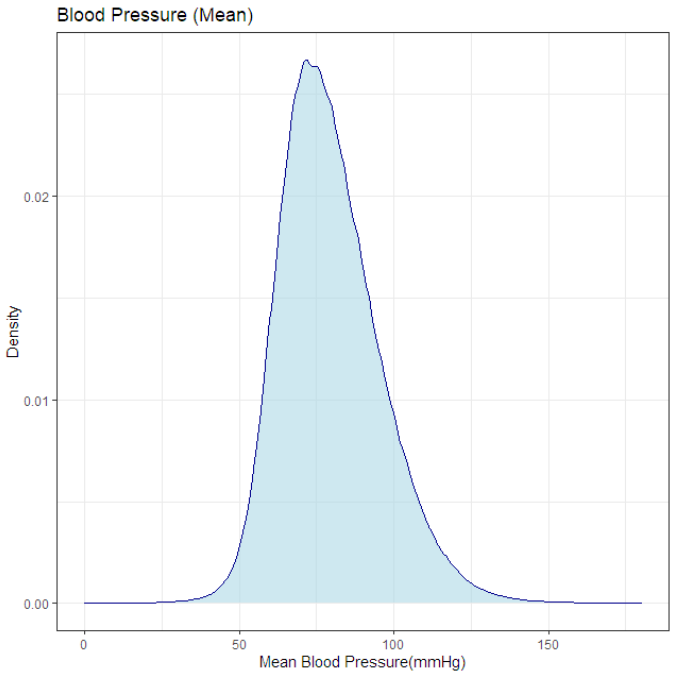


Figure 5-7 –Heart Rate distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right)

Blood Pressure (Mean)



Blood pressure is the force of the blood pushing against the artery walls during contraction and relaxation of the heart. Two numbers are recorded when measuring blood pressure. The higher number, or systolic pressure, refers to the pressure inside the artery when the heart contracts and pumps blood through the body. The lower number, or diastolic pressure, refers to the pressure inside the artery when the heart is at rest and is filling with blood. Low Mean Arterial Pressure (MAP) has been correlated with higher risk of mortality (Maheshwari et al., 2018).

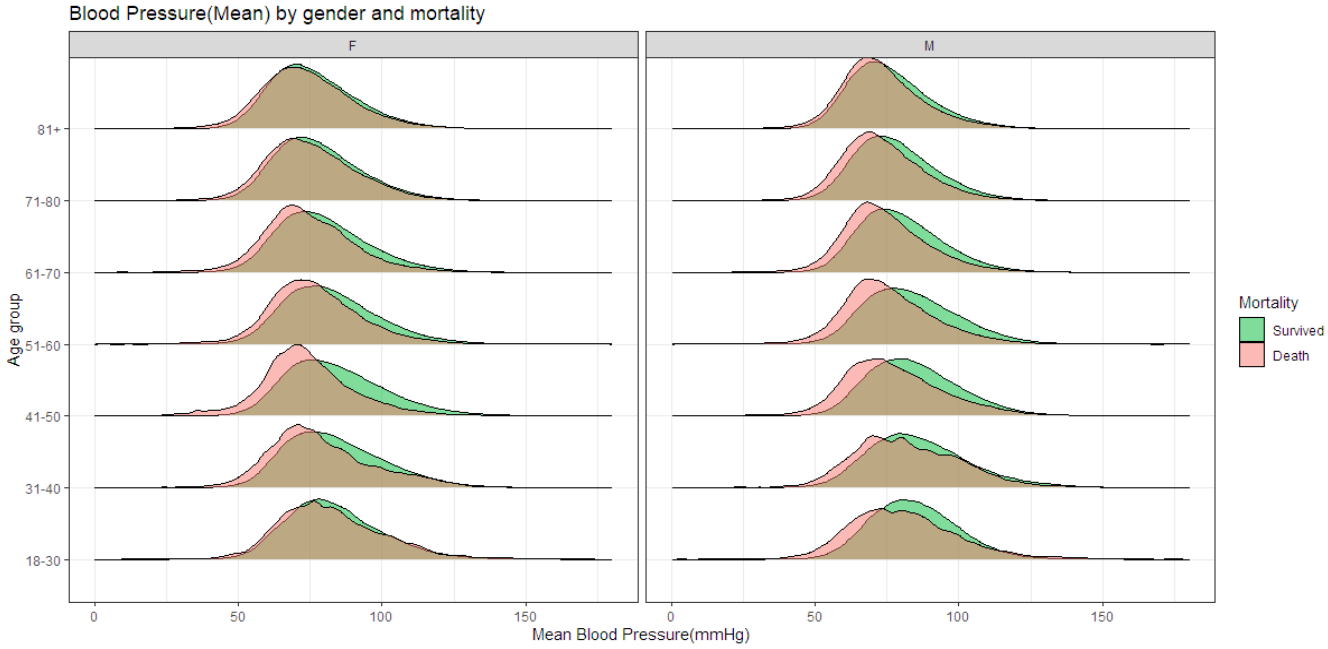
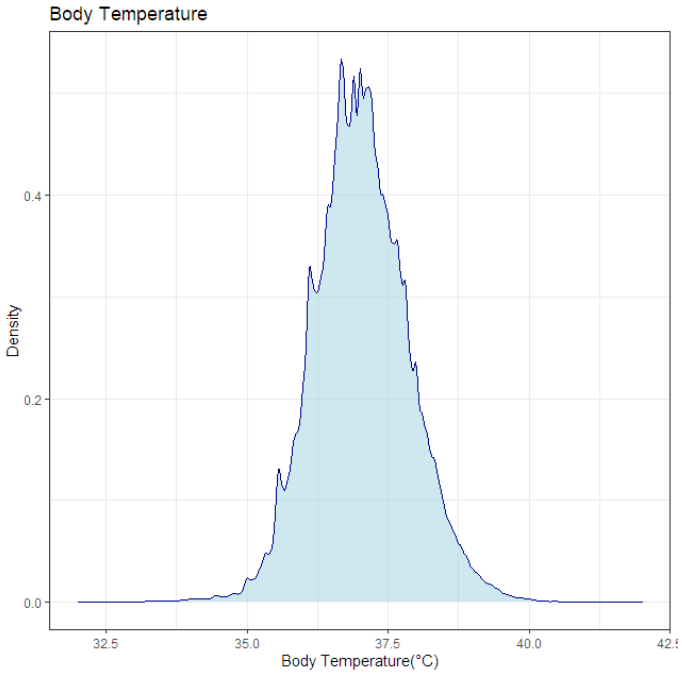


Figure 5-8 – Mean Blood Pressure distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right)

Body Temperature



The normal body temperature of a person varies depending on gender, recent activity, food and fluid consumption, time of day, and, in women, the stage of the menstrual cycle. Normal body temperature can range from 36.5 degrees C, to 37.2 degrees C for a healthy adult. In the graph below we can see that hypothermia conditions are usually more dangerous than hyperthermia.

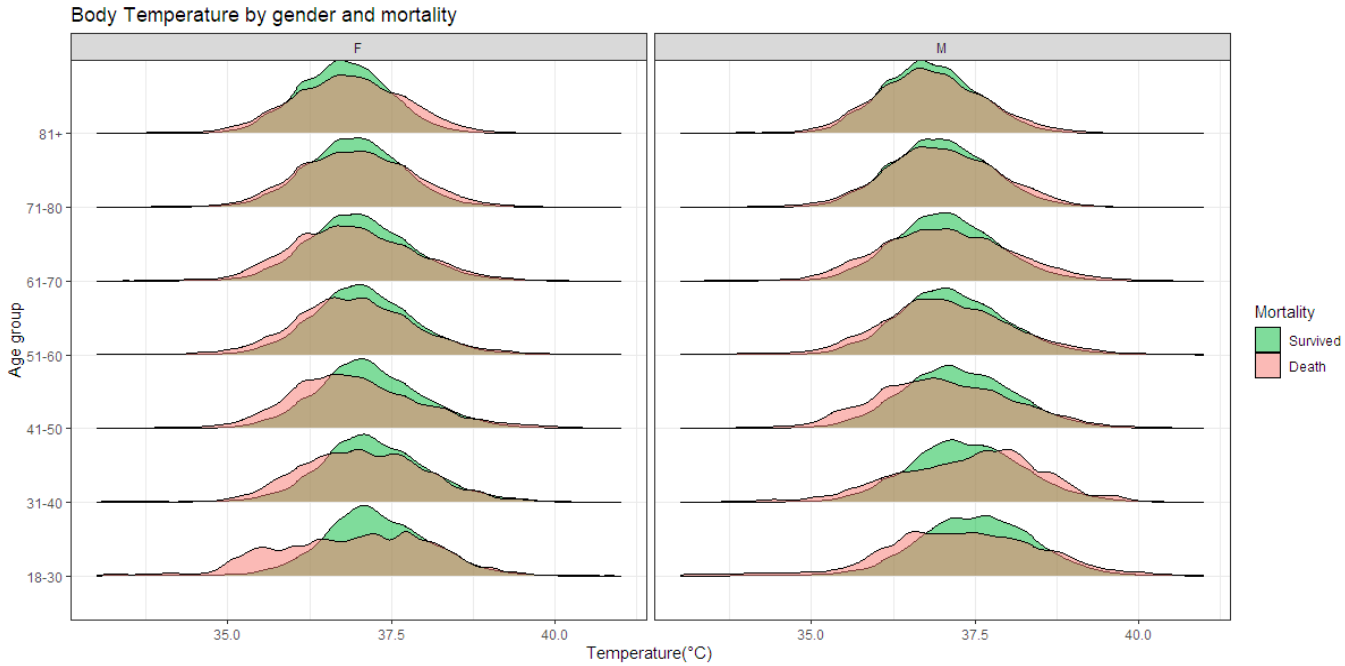
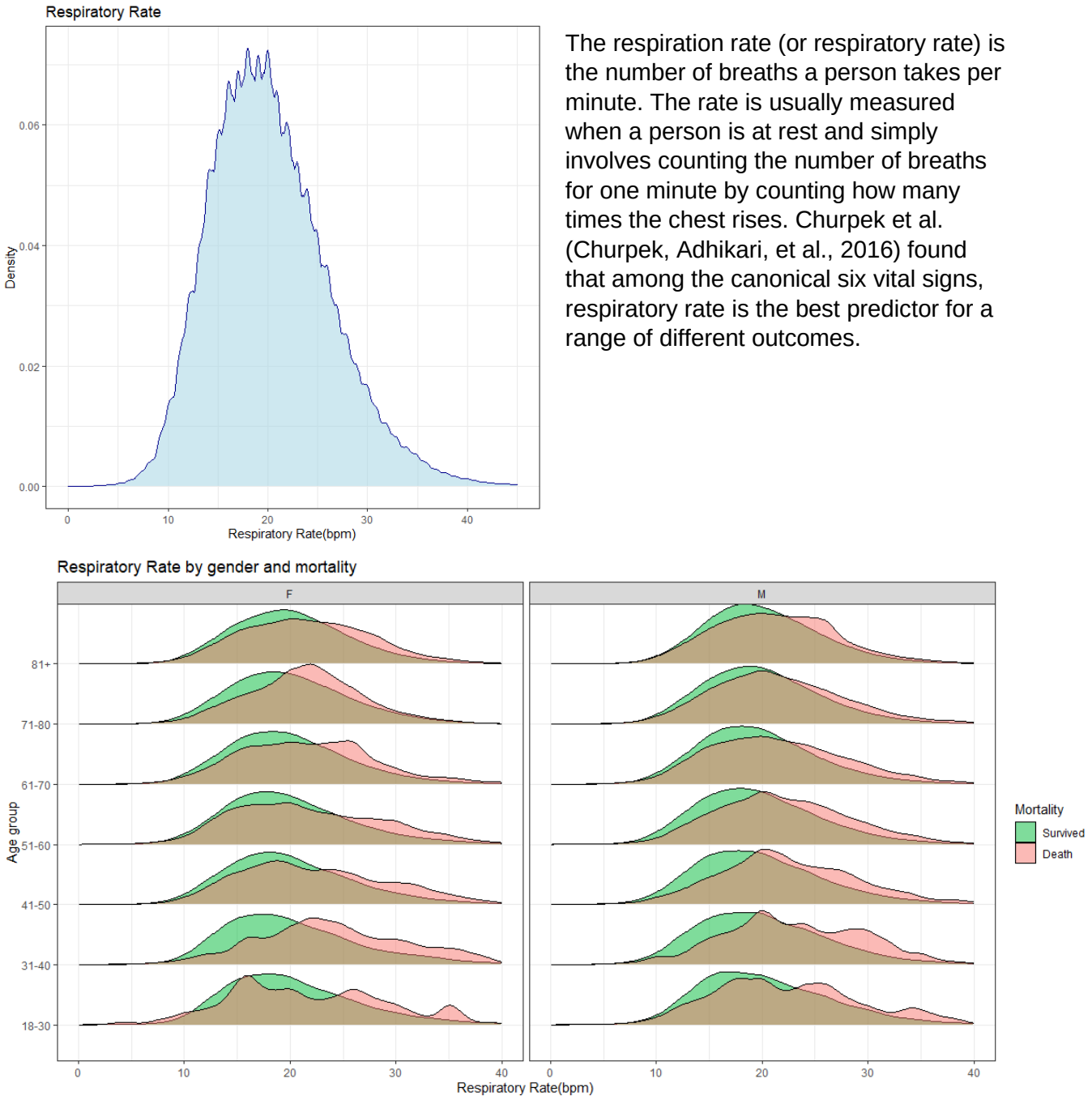


Figure 5-9 – Body Temperature (°C) distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right)

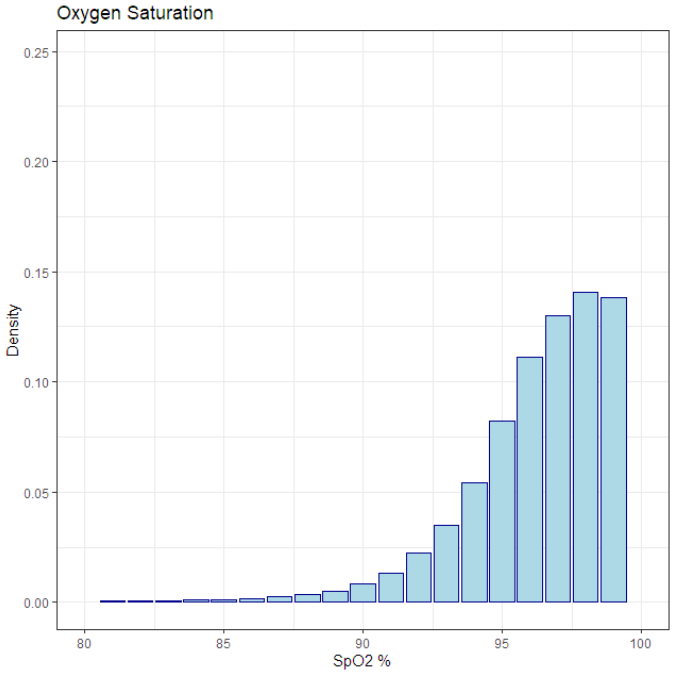
Respiration Rate



The respiration rate (or respiratory rate) is the number of breaths a person takes per minute. The rate is usually measured when a person is at rest and simply involves counting the number of breaths for one minute by counting how many times the chest rises. Churpek et al. (Churpek, Adhikari, et al., 2016) found that among the canonical six vital signs, respiratory rate is the best predictor for a range of different outcomes.

Figure 5-10 – Respiratory Rate distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right)

Oxygen Saturation



SpO2, also known as oxygen saturation, is a measure of the amount of oxygen-carrying haemoglobin in the blood relative to the amount of haemoglobin not carrying oxygen. Oxygen saturation levels below 94% have shown to correlate with higher risk of mortality in ICU (Zhou et al., 2020).

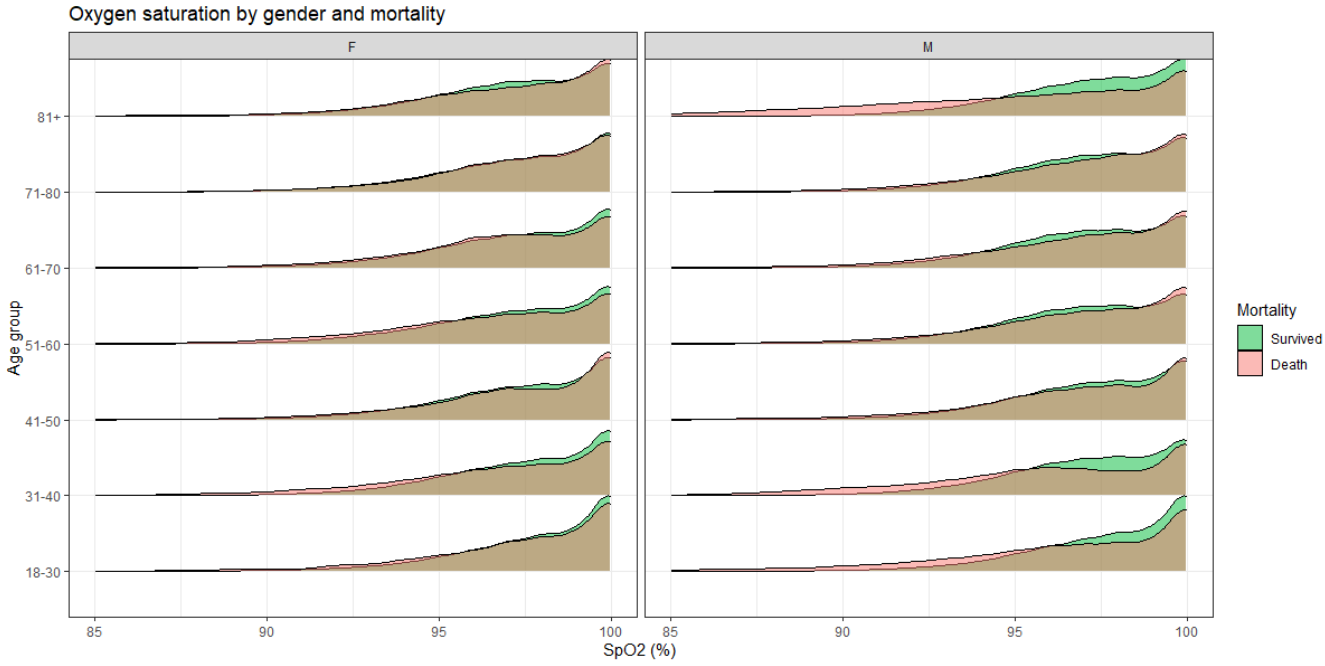
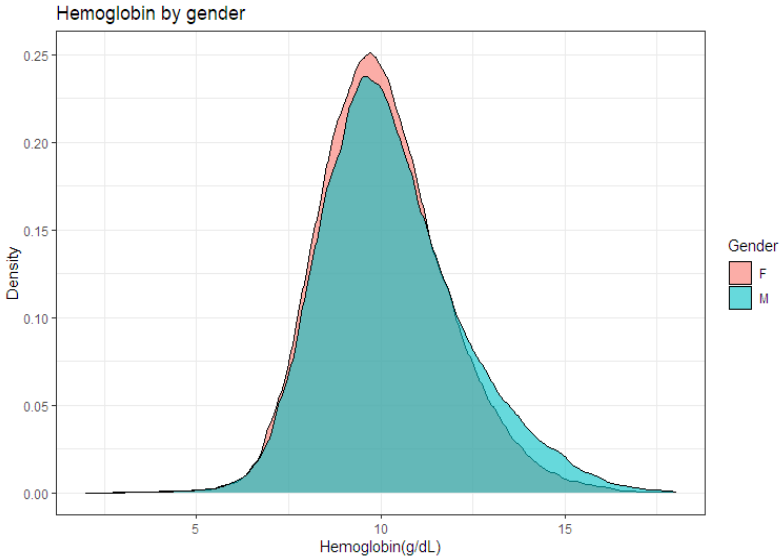


Figure 5-11 – Oxygen Saturation distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right)

5.4.6 Laboratory results

The only difference in layout between VS and Laboratory results is that the distribution in the top left corner takes gender differences into account.

Hemoglobin



Hemoglobin is a protein in red blood cells that carries iron. This iron holds oxygen, making haemoglobin an essential component of the blood. When the blood doesn't contain enough haemoglobin, cells don't receive enough oxygen.

Hemoglobin levels below 9g/dl have been correlated with higher risk of mortality (Jung et al., 2019).

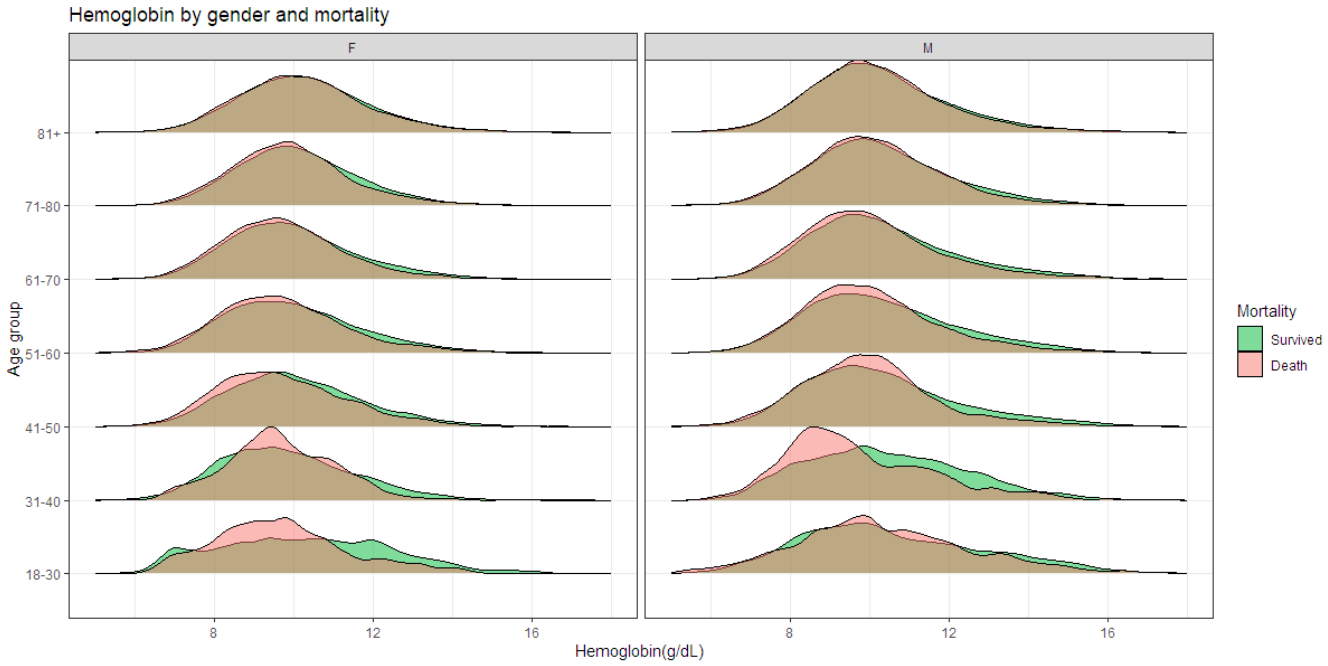
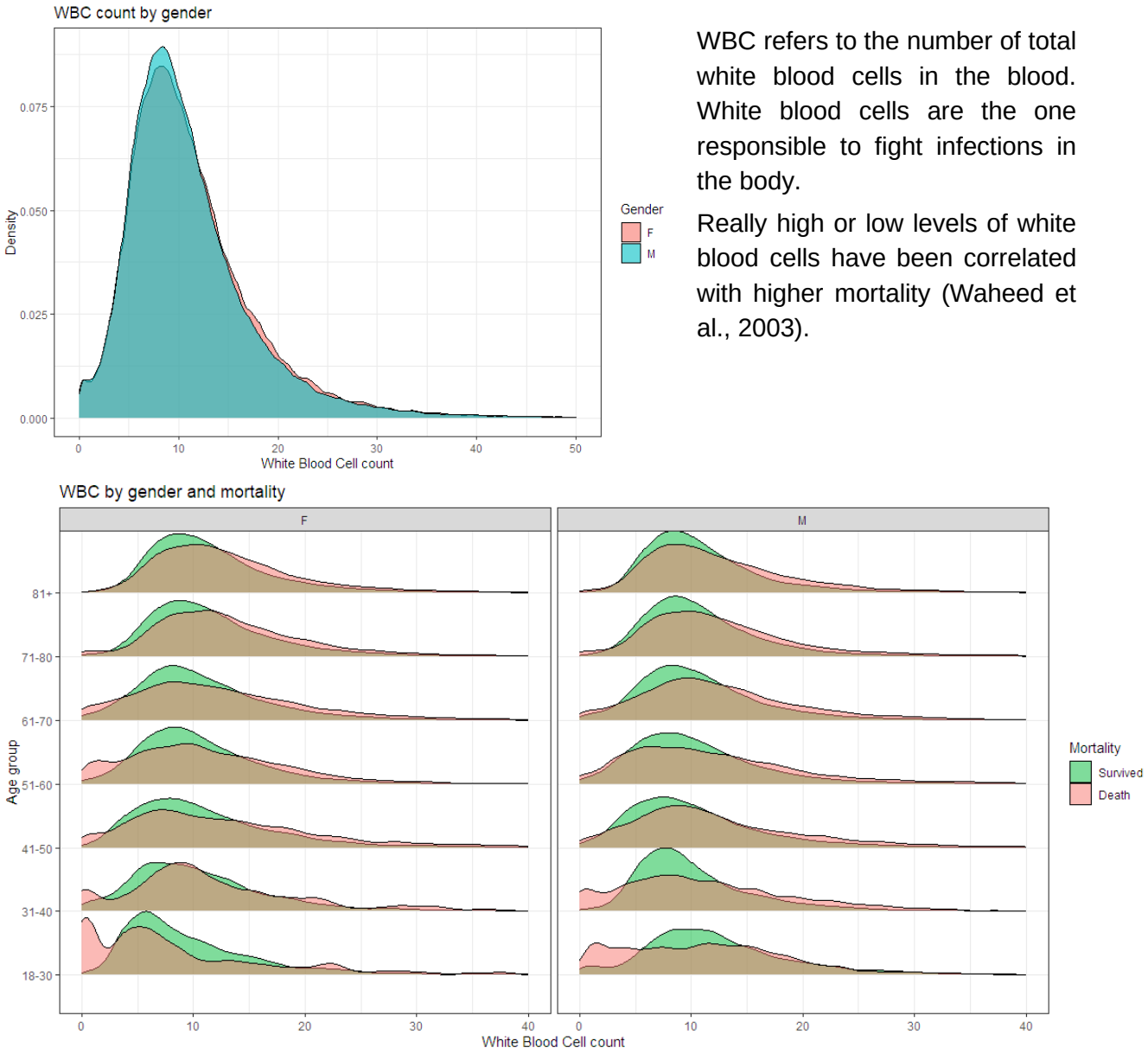


Figure 5-12 – Haemoglobin distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right)

White Blood Cells Count

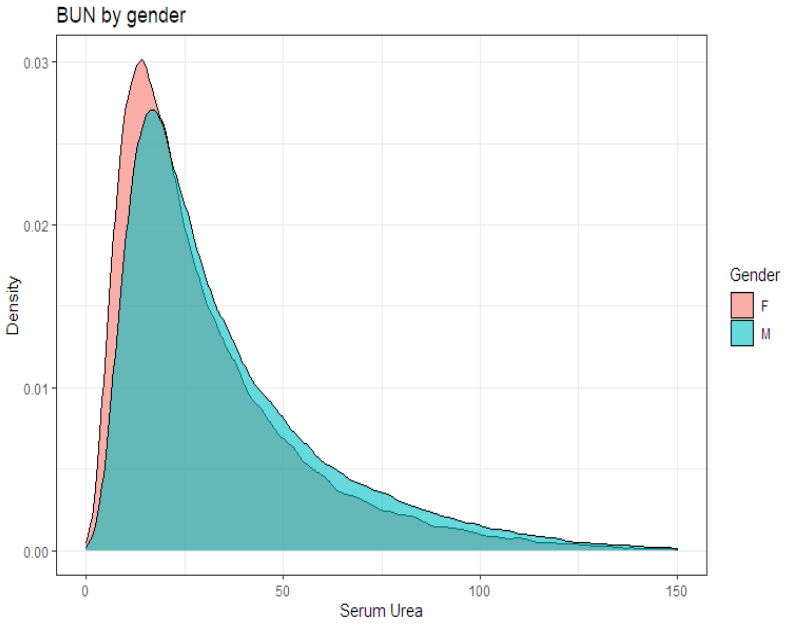


WBC refers to the number of total white blood cells in the blood. White blood cells are the one responsible to fight infections in the body.

Really high or low levels of white blood cells have been correlated with higher mortality (Waheed et al., 2003).

Figure 5-13 – White Blood Cell count distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right)

Blood Urea Nitrogen



A blood urea nitrogen (BUN) test is used to determine how well your kidneys are working. It does this by measuring the amount of urea nitrogen in the blood. Urea nitrogen is a waste product that's created in the liver when the body breaks down proteins. People with really high BUN levels (over 75) often need dialysis, and extremely high BUN levels (over 200) have been correlated with high mortality risk (26% in (do Nascimento et al., 2010)) .

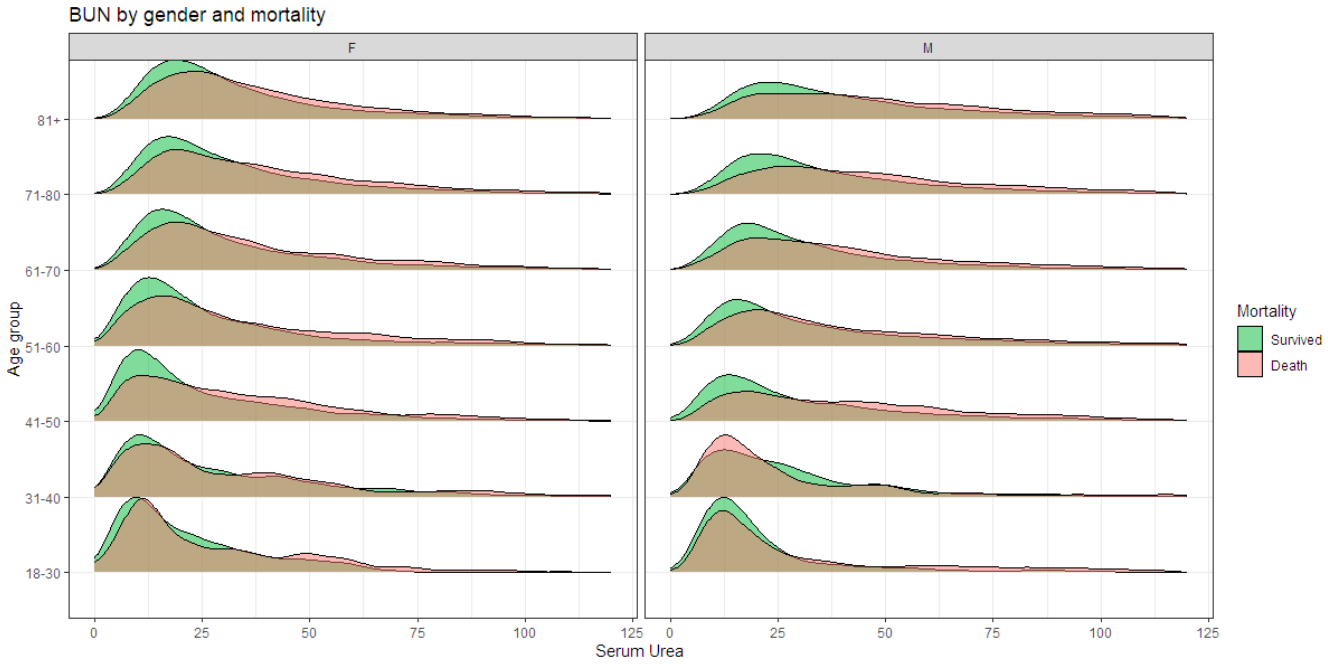
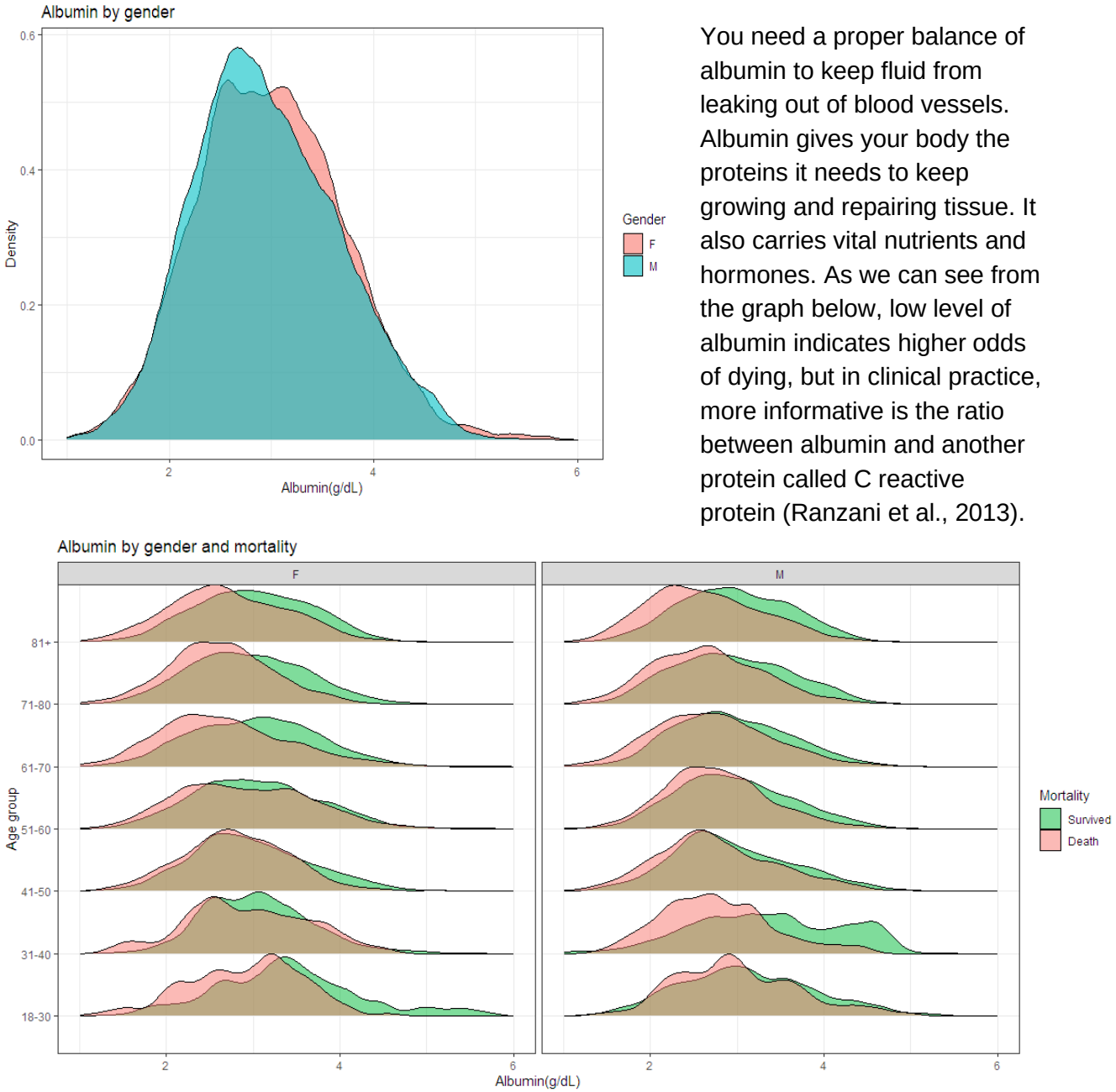


Figure 5-14 – Blood Urea Nitrogen distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right)

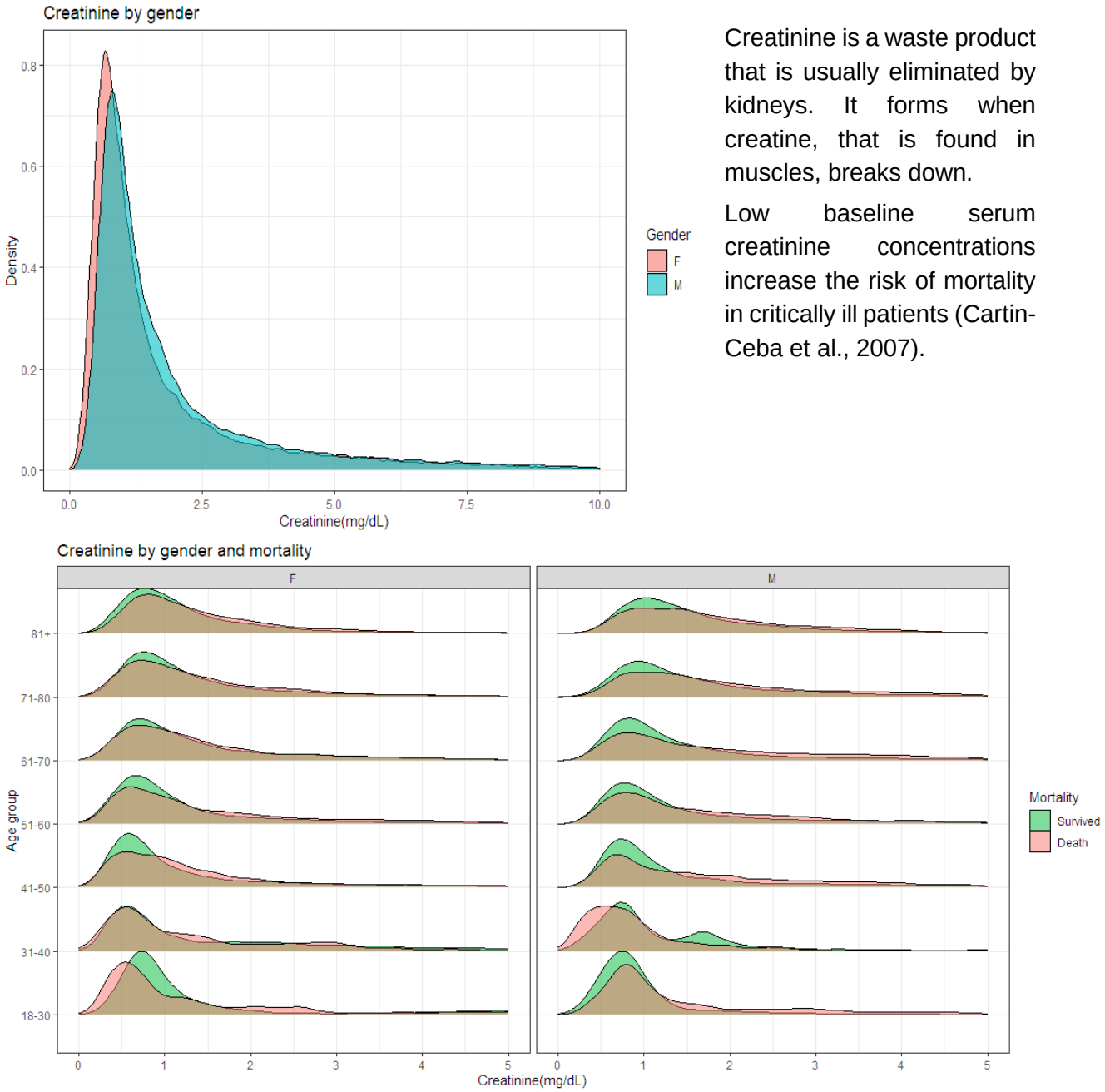
Albumin



You need a proper balance of albumin to keep fluid from leaking out of blood vessels. Albumin gives your body the proteins it needs to keep growing and repairing tissue. It also carries vital nutrients and hormones. As we can see from the graph below, low level of albumin indicates higher odds of dying, but in clinical practice, more informative is the ratio between albumin and another protein called C reactive protein (Ranzani et al., 2013).

Figure 5-15 – Albumin distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right)

Creatinine

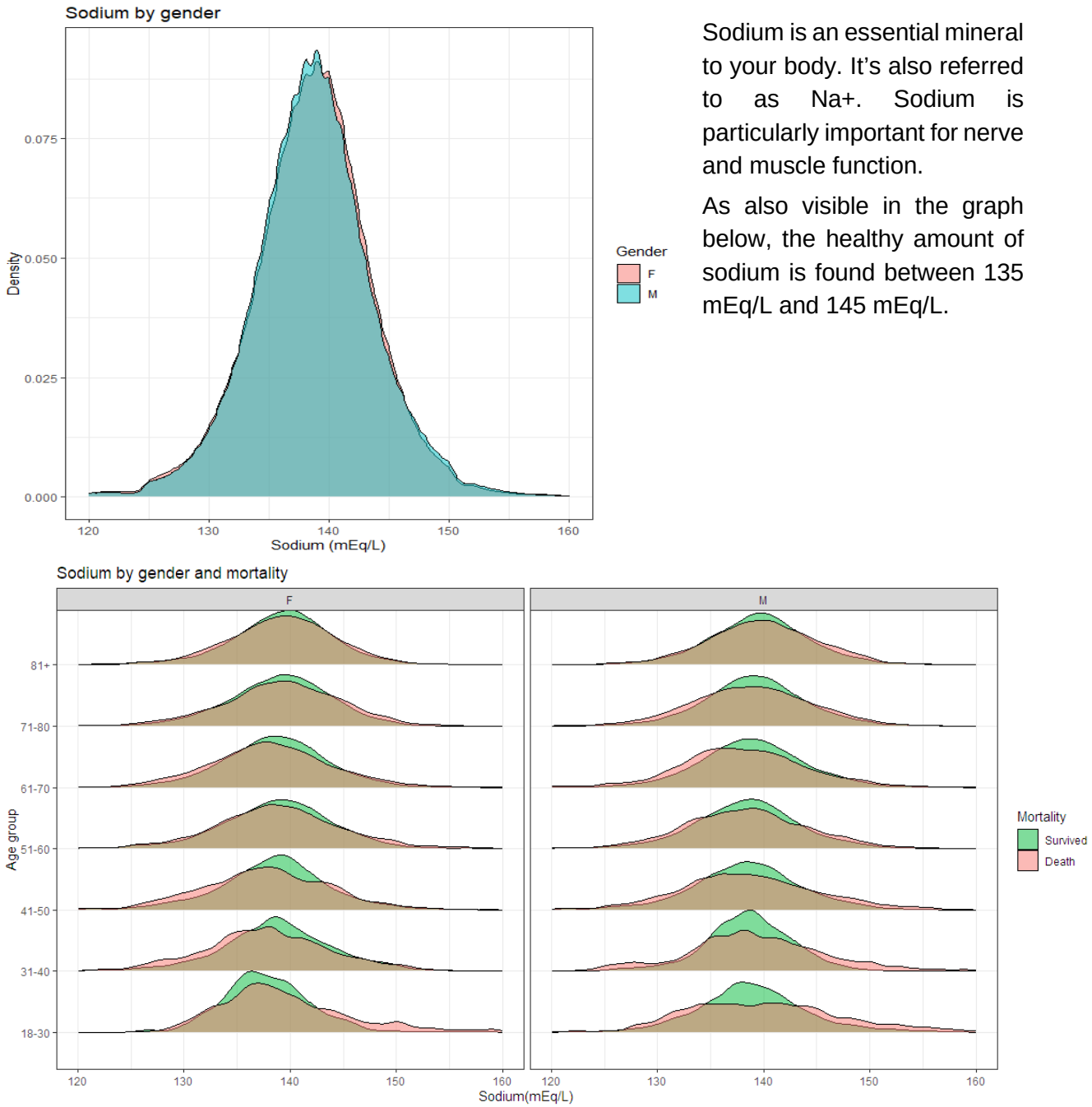


Creatinine is a waste product that is usually eliminated by kidneys. It forms when creatine, that is found in muscles, breaks down.

Low baseline serum creatinine concentrations increase the risk of mortality in critically ill patients (Cartin-Ceba et al., 2007).

Figure 5-16 – Creatinine distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right)

Sodium

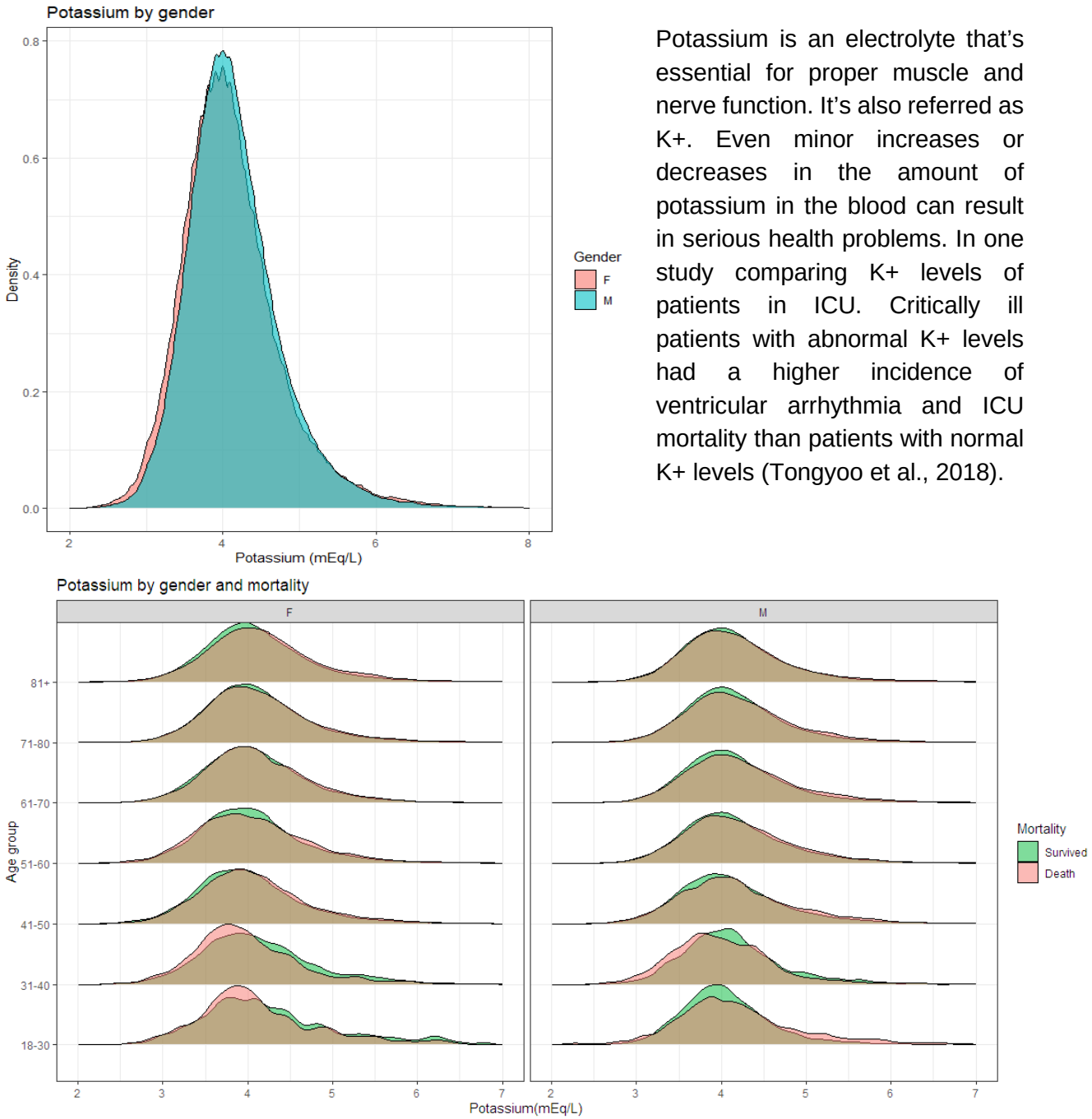


Sodium is an essential mineral to your body. It's also referred to as Na+. Sodium is particularly important for nerve and muscle function.

As also visible in the graph below, the healthy amount of sodium is found between 135 mEq/L and 145 mEq/L.

Figure 5-17 – Sodium distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right)

Potassium



Potassium is an electrolyte that's essential for proper muscle and nerve function. It's also referred as K+. Even minor increases or decreases in the amount of potassium in the blood can result in serious health problems. In one study comparing K+ levels of patients in ICU. Critically ill patients with abnormal K+ levels had a higher incidence of ventricular arrhythmia and ICU mortality than patients with normal K+ levels (Tongyoo et al., 2018).

Figure 5-18 – Potassium distribution for the whole cohort (top left corner). Ridge graphs of mortality (colour coded) broken down by age group (y axis) and gender (Females on the left and Males on the right)

Since Albumin recordings are rare (less than 10% of the rows in the lab results have albumin recordings), one interesting thing was to understand if people that undergo an albumin test are more likely to die:

```
> observed_albumin
survived_perc    death_perc
        0.892         0.108
> observed_noalbumin
survived_perc    death_perc
        0.945         0.055
```

Subjects that have an albumin recording have 5% higher mortality rate.

Another interesting insight is to see if patients with more frequent recordings are more likely to die (i.e. are checked more frequently due to wore health condition). That was the case with all the laboratory results.
In the following code snippet we can see the example of Hemoglobin recordings.

gt1000 stands for greater than 1000 (minutes between successive recordings). lt1000 stands for less than 1000.

```
> observed_gt1000
survived_perc    death_perc
        0.924         0.076
> observed_lt1000
survived_perc    death_perc
        0.887         0.113
```

Subject with more frequent recordings (every 8h or 16h vs. every 24h or more) of Hemoglobin have almost 4% higher mortality.

6 Data Pre-processing

6.1 The problem of time intervals

When dealing with time series everything gets really messy and complicated if they don't have the same sample rate. In the real world that's almost always the case, unless you are the one that records the data in the first place and applies a strict protocol. For hospital recorded data, of course, there's no precise and unique timing for the measurements, and that's why the problem of discretizing time intervals arises. Most AI models need fixed shape inputs and circumvent this constraint is usually a complex task.

For example, suppose that we want to predict mortality using 24 hours' worth of data. The first thing that we need to do is establish a starting point: for this study, it will be the time of the first vital sign recording in ICU. After setting the starting point, cutting 24 hours of data will return time series of varying duration, because some patients will have more recordings than others. To solve this problem and obtain fixed-length 24h time series, we have to decide how many time points we want in a 24h period, i.e. setting the length of the time intervals. Let's assume for now to have set the time window length to *t*, at this point we will have to face another problem: what to do when we find multiple data points in the same window. We have to set a policy to aggregate them into a single value (median value has been used in this project). Also, when no point is found in a time window another kind of policy must be endorsed to deal with the missing value. When just a single data point is found in a time window no problem arises. Those policies will be discussed when dealing with missing data.

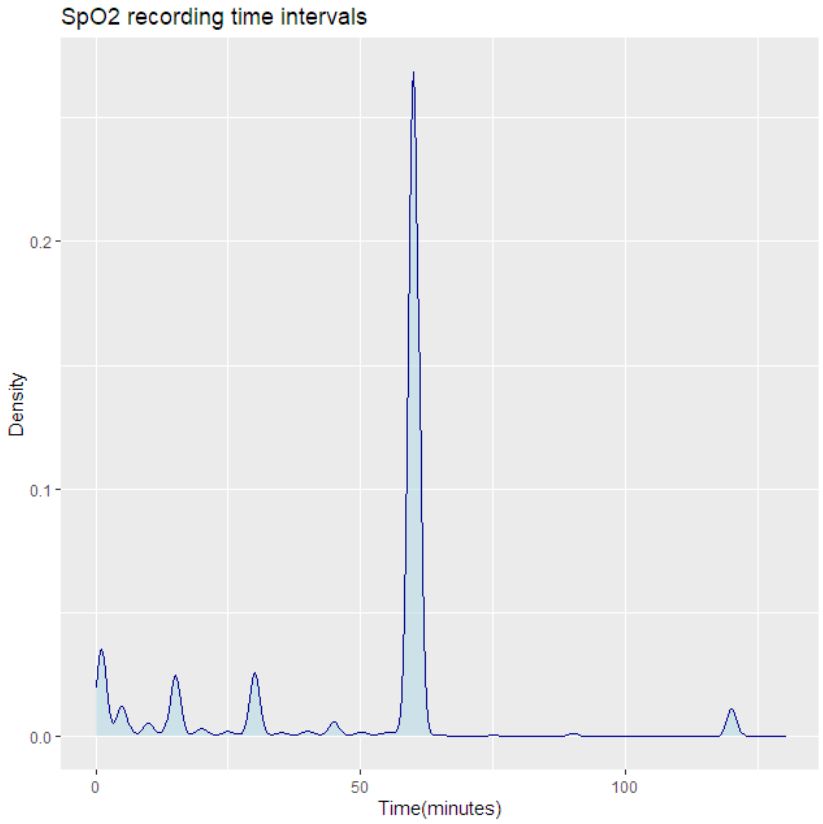
Now, going back to the choice of the length t of the time window, if we choose one that is too wide we will lose information by doing frequent aggregation, otherwise, if we choose one that is too narrow, we will add a lot of empty points that need to be filled with artificial data. A clever way to set the ideal size of the time window is to look at the distributions of time intervals in the recordings hoping to find a common pattern.

With a visual inspection of the below graphs we can deduce that most *vital signs* show a main peak at 1h interval. Glucose and Temperature have a different behaviour. After the inspection, to uniform the vital signs data, the discretization interval has been set to 1h even if glucose and temperature show different behaviours.

Glucose measurements are done more erratically (it is not a routine measurement) but they also show the main peak at 1h. Temperature have multiple hourly peaks, with the main peak at 4h interval, as also suggested by the distribution of measurements during the day.

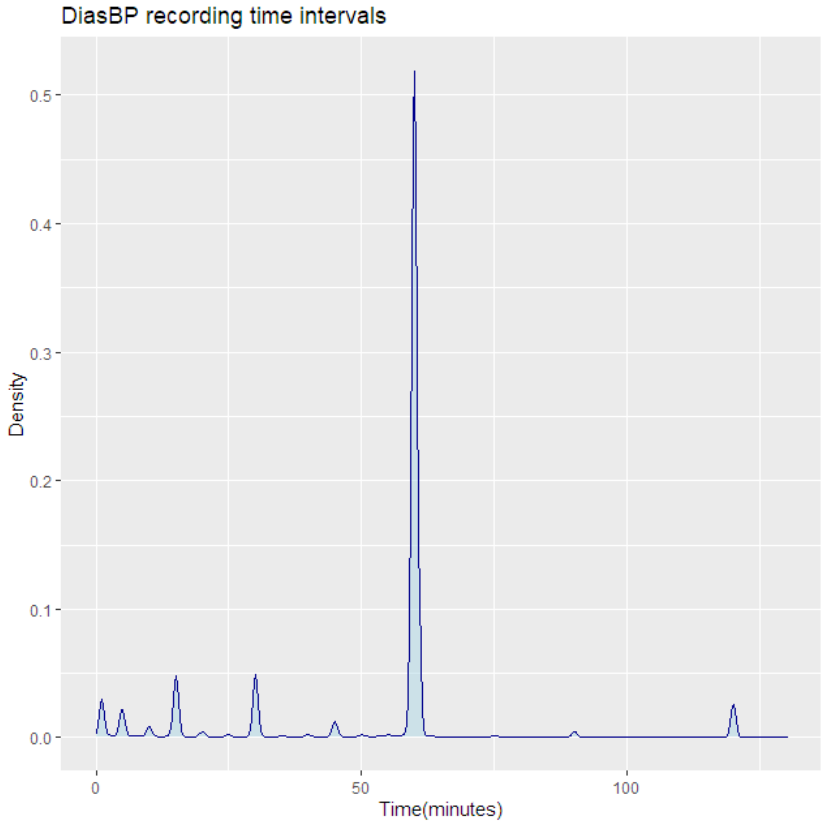
The summary of the distributions have a really high max value: this is due to a few number of outliers that had multiple ICU stays recorded with the same ID.

For what concerns *laboratory results* intervals, all of them have similar distributions: two small peaks at around 8h and 14h interval and then a bigger peak at 24h interval. With this in mind, the best compromise for a good time interval is 8h.



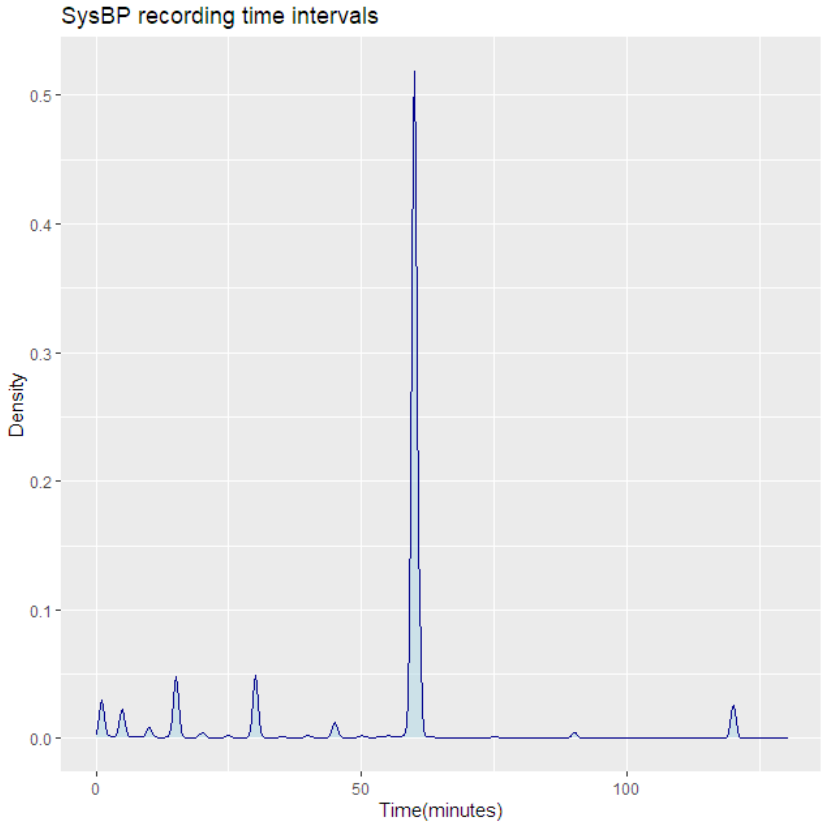

```
Min.    :    1
1st Qu.:   30
Median :   60
Mean   :   51
3rd Qu.:   60
Max.   :37023
```

Figure 6-1 – SpO2 recording time intervals with summary statistics



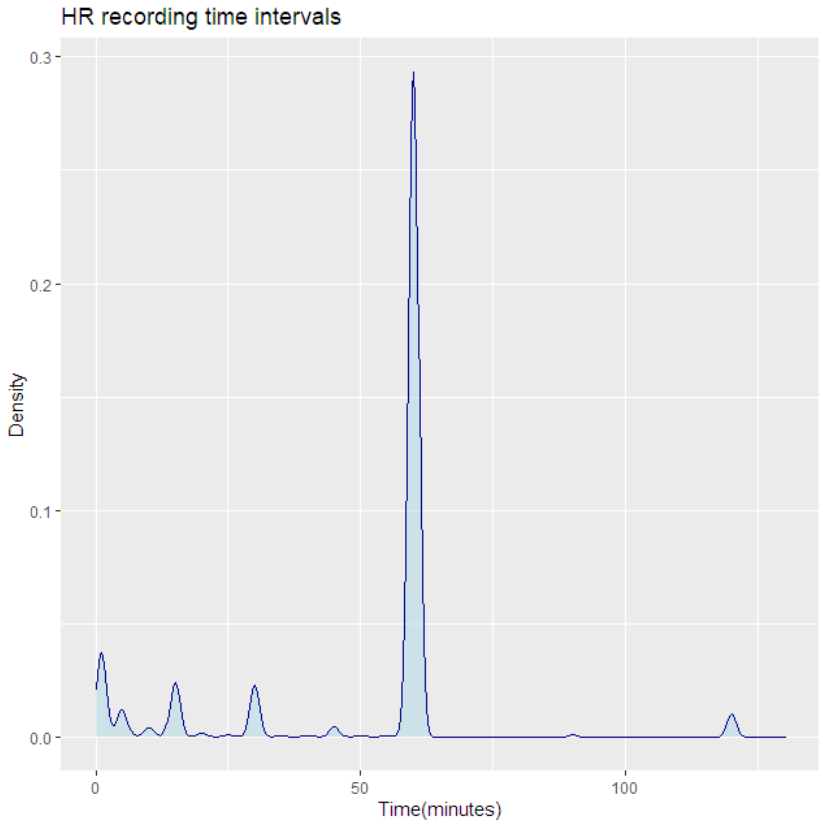

```
Min.    :    1.00
1st Qu.:   46.00
Median :   60.00
Mean   :   54.04
3rd Qu.:   60.00
Max.   :33780.00
```

Figure 6-2 – Diastolic Blood Pressure recording time intervals with summary statistics



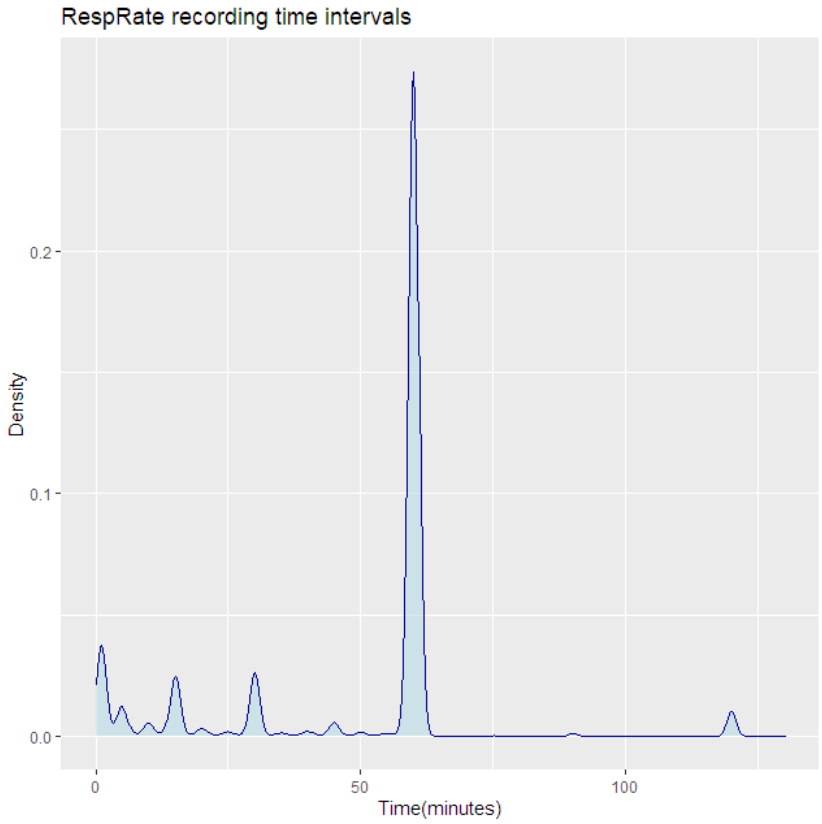

```
Min.    :    1.00
1st Qu.:   46.00
Median :   60.00
Mean   :   54.04
3rd Qu.:   60.00
Max.   :33780.00
```

Figure 6-3 – Systolic Blood Pressure recording time intervals with summary statistics

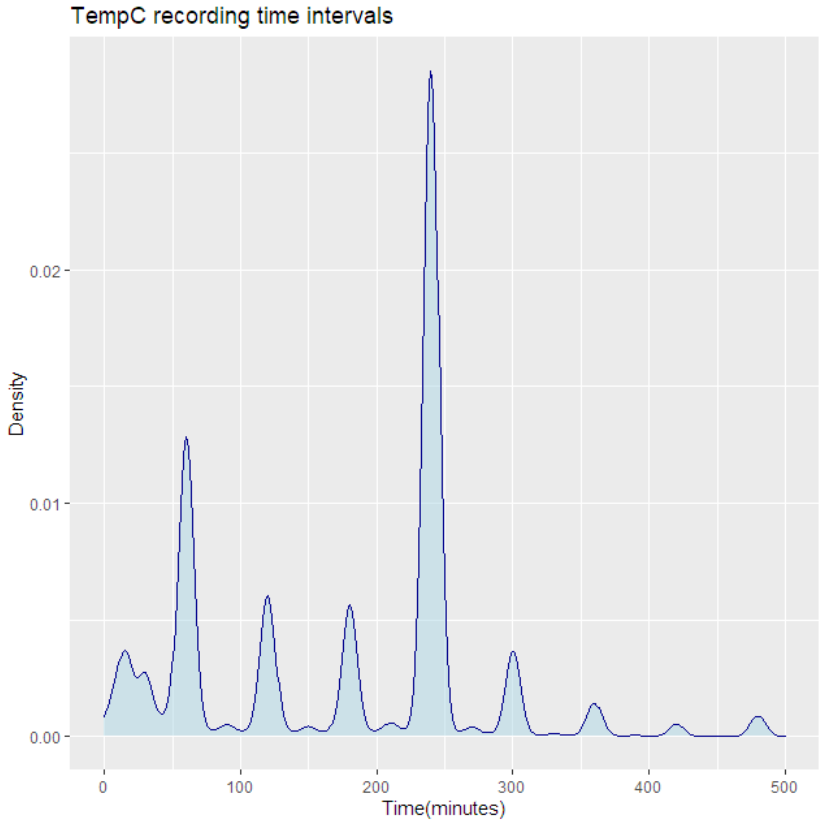



```
Min.    :    1.00
1st Qu.:   46.00
Median :   60.00
Mean   :   54.04
3rd Qu.:   60.00
Max.   :33780.00
```

Figure 6-4 – Heart Rate recording time intervals with summary statistics

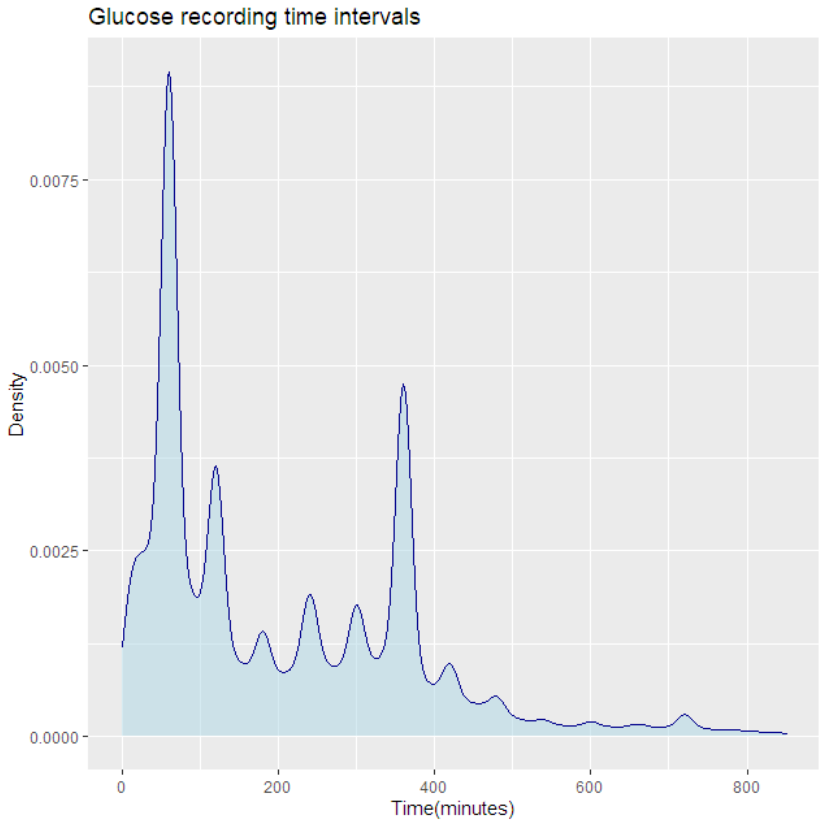



```
Min.    :    1.00
1st Qu.:   30.00
Median :   60.00
Mean   :   49.67
3rd Qu.:   60.00
Max.   :33780.00
```



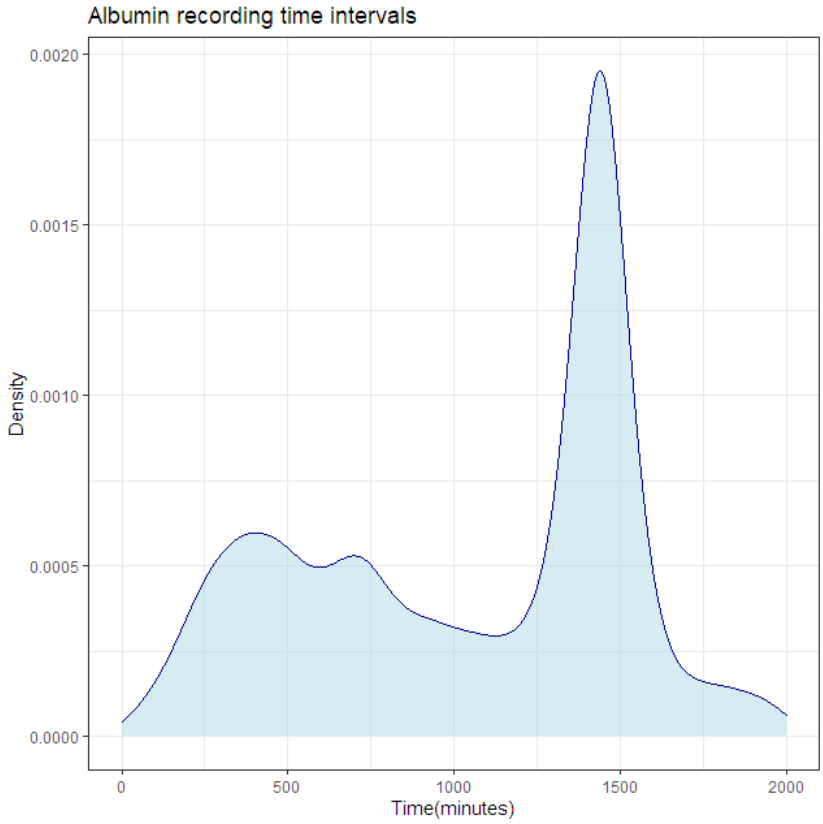

```
Min.    :    1.0
1st Qu.:   60.0
Median :  240.0
Mean   :  181.3
3rd Qu.:  240.0
Max.   :33900.0
```

Figure 6-5 – Body Temperature (°C) recording time intervals with summary statistics



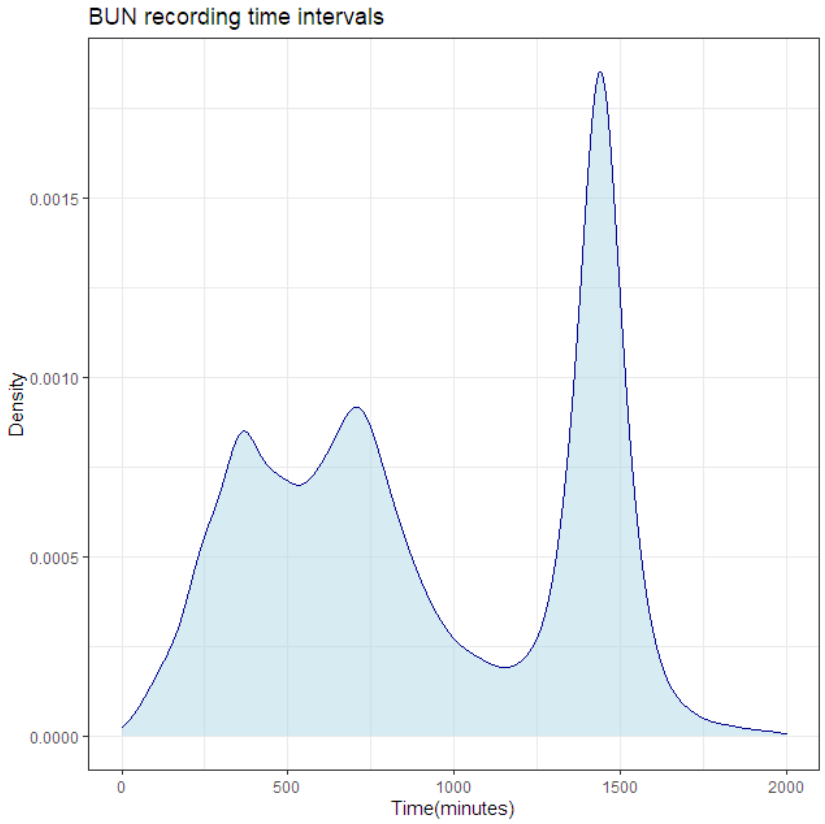

```
Min.    :    1.0
1st Qu.:   60.0
Median :  155.0
Mean   :  237.6
3rd Qu.:  360.0
Max.   :33840.0
```

Figure 6-6 – Glucose recording time intervals with summary statistics



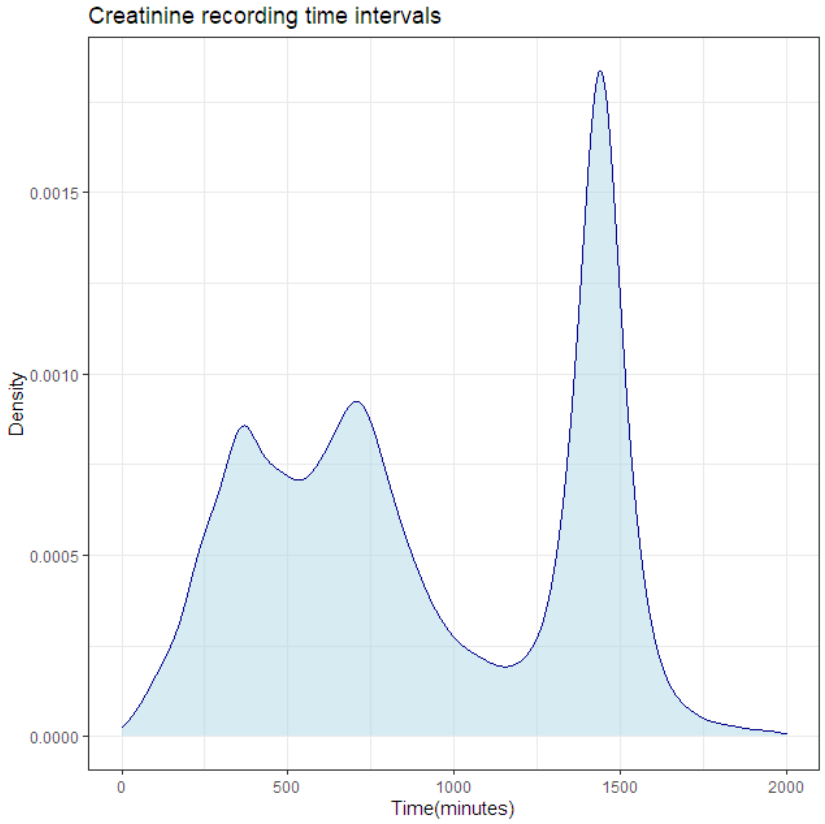

```
Min.    :    1.0
1st Qu.:  919.8
Median : 1460.0
Mean   : 2878.7
3rd Qu.: 2997.0
Max.   :72972.0
```

Figure 6-7 – Albumin recording time intervals with summary statistics



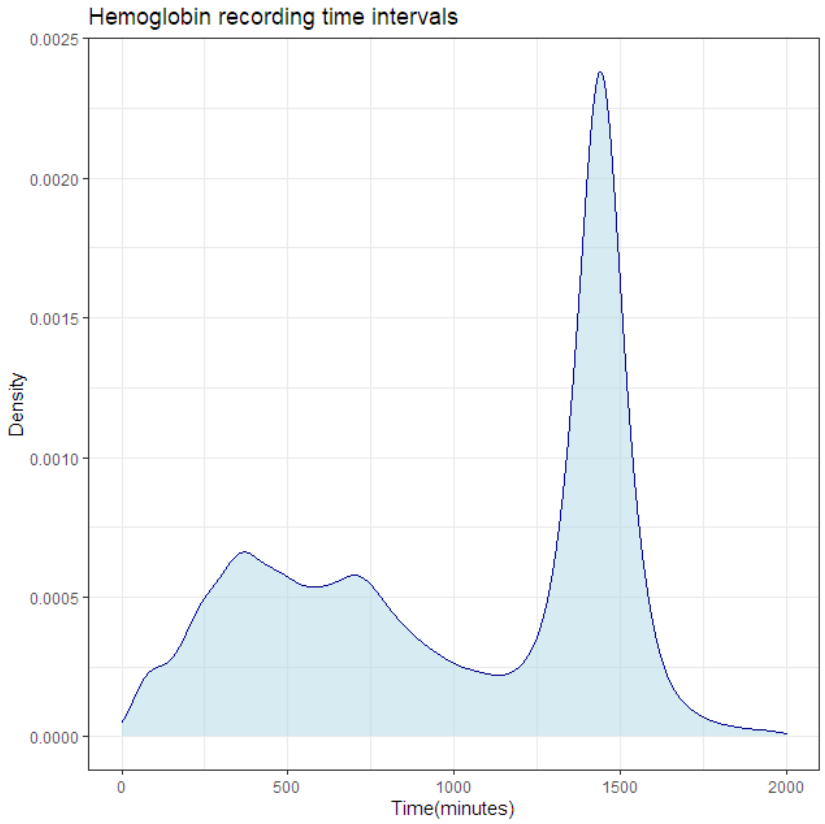

```
Min.    :    1
1st Qu.:  508
Median :  826
Mean   :  946
3rd Qu.: 1416
Max.   :95400
```

Figure 6-8 – Blood Urea Nitrogen recording time intervals with summary statistics



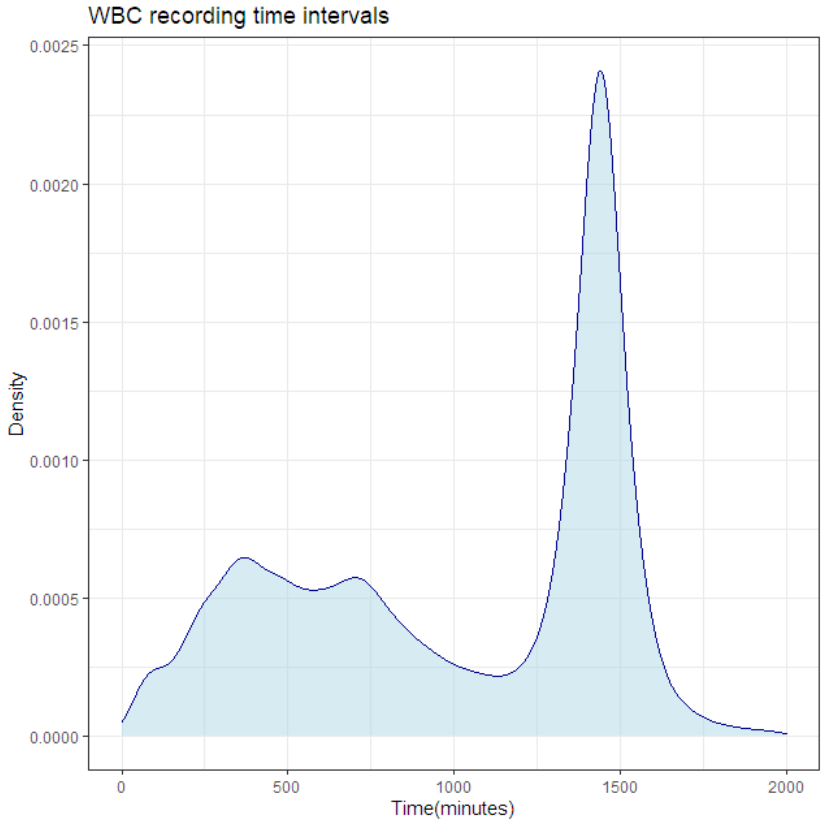

```
Min.    :    1.0
1st Qu.:  505.0
Median :  820.0
Mean   :  942.2
3rd Qu.: 1415.0
Max.   :95400.0
```

Figure 6-9 – Creatinine recording time intervals with summary statistics



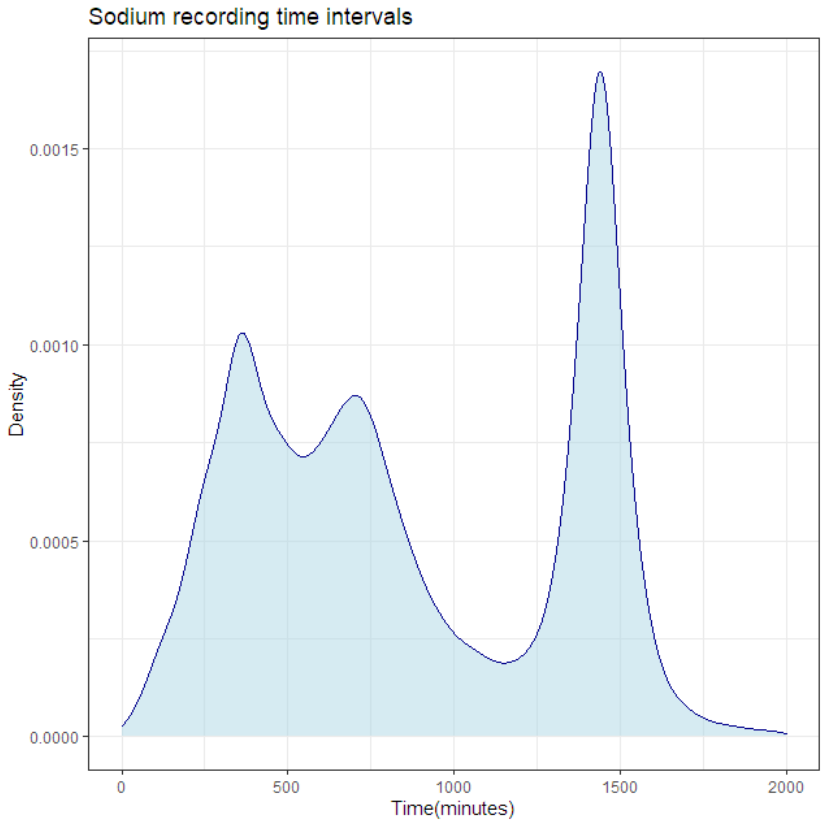

```
Min.    :     1
1st Qu.:   563
Median :  1239
Mean   :  1061
3rd Qu.:  1444
Max.   :162080
```

Figure 6-10 – Hemoglobin recording time intervals with summary statistics



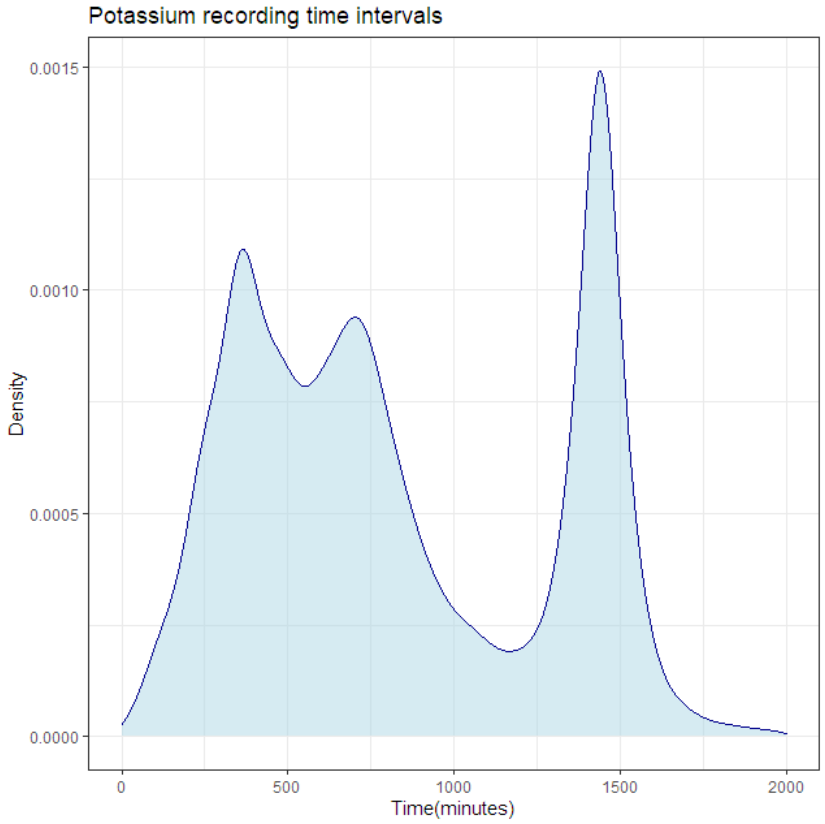

```
Min.    :     1
1st Qu.:   570
Median :  1253
Mean   :  1064
3rd Qu.:  1444
Max.   :198675
```

Figure 6-11 – White Blood Cell count recording time intervals with summary statistics




```
Min.    :    1.0
1st Qu.:  454.0
Median :  777.0
Mean   :  916.1
3rd Qu.: 1405.0
Max.   :95400.0
```

Figure 6-12 – Sodium recording time intervals with summary statistics




```
Min.    :    1.0
1st Qu.:  440.0
Median :  735.0
Mean   :  877.3
3rd Qu.: 1376.0
Max.   :95400.0
```

Figure 6-13 – Potassium recording time intervals with summary statistics

6.2 Dealing with outliers

Three main types of outlier detection techniques exist:

1. domain knowledge-based outlier removal (by removing impossible and/or implausible values when the possible range of values is well known)
2. statistical based outlier removal (Inter Quartile Range (IQR) based, Generalized Extreme Studentized Deviate test (GESD), Grubbs etc.)
3. Machine learning based outlier removal (density clustering approaches, elimination forests, one class Support Vector Machines (SVM) etc.).

Gaspar et al. (Gaspar et al., 2011), in 2011, found that most research that uses medical data applies statistical methods for outlier's detection. Moreover, in the work of Maslove et al. (Maslove et al., 2016) outlier detection methods specifically for MIMIC-III have been studied: first, they removed logical outlier (highly implausible values like non-terminal zeroes or impossible BP values), then they divided data in gaussian distributed and non-gaussian distributed. Grubb's and ESD tests were used on gaussian data and IQR based outlier detection on non-gaussian ones.

Keeping this in mind, outlier detection and removal have been implemented in two steps:

- **First step** – Looking at data's boxplots. This led to the understanding that not every distribution of data had a natural range of values (e.g. glucose had recordings over 1000 and temperature had recordings at less than 32°C). Some gross outlier could be removed by visual inspection of those boxplots. Then literature was searched for plausible ranges of values (Table 6-1) to reduce them accordingly.
- **Second step** – As reported in (Maslove et al., 2016), values with really high odds of being outliers have been removed: nonterminal zeroes and recordings with impossible BP values (Sys<Dias, Sys<Mean, Mean<Dias).

Feature	Min value	Max value
Heart Rate	0	300
Blood Pressure (Systolic)	0	560
Blood Pressure (Diastolic)	0	360
Respiratory Rate	0	70
Temperature (°C)	30	43
Oxygen Saturation (SpO2)	50	100
Glucose	0	600
Hemoglobin	0	30
White Blood Cell count (WBC)	0	1000
Blood Urea Nitrogen (BUN)	0	250
Albumin	0	20
Creatinine	0	50
Sodium	0	580
Potassium	0	30

Table 6-1 - Plausible extreme value ranges

After those two steps, Generalized Extreme Studentized Deviate test (GESD) on normally distributed data (checked with Shapiro-Wilk test) and IQR based outlier detection on non-normal data were applied. Even if the more extreme values detected with these methods seem anomalous values, many other extreme values could be physiological and, as a consequence of the application of these techniques, the high entropy data could be discarded in favour of the elimination of few anomalous values. Therefore, the decision was made to not apply statistical techniques and keep data as raw as possible.

Because machine learning approaches to the detection of outliers can be really time consuming, I decided to not delve into those techniques.

Observations	**# of observations**	**% on the whole dataset**
Total Vital Signs data	5736975	
Invalid Blood Pressure	12917	(0.225%)
Nonterminal zeroes	30	(0.001%)
Implausible Temperature	56	(0.001%)
Implausible SpO2	2255	(0.039%)
Total Vital Signs' Outliers	15224	(0.265%)

Table 6-2 - Number of outliers (Vital Signs)

Observations	**# of observations**	**% on the whole dataset**
Total Laboratory Results data	863744	
Implausible BUN	18	(0.002%)
Implausible Creatinine	3	(0.000%)
Total Laboratory Results' Outliers	21	(0.002%)

Table 6-3 - Number of outliers' data (Laboratory Results)

6.3 Dealing with missing values

As a result of the time discretization and outlier removal steps, but also as an intrinsic characteristic of clinical data, a high number of missing values are scattered all over the dataset.

The simplest approaches to follow when dealing with missing values would be elimination or filling with a special value (usually 0 or -1). Those approaches can work well when the percentage of missing values over the whole dataset is relatively small and, unfortunately, that's not the case.

Two different techniques have been implemented, leaving two different copies of the dataset for the classification tests:

1. Filling missing data with previous or successive values. To leave no blanks, for vital signs data a simple carry forward filling and a successive carry backward filling have been applied. For laboratory data a slightly different approach has been followed. Since laboratory results trends change at a slower pace than vital signs (and have also been discretized to 8h interval of recordings), is reasonable, when the first value of the series is missing, to search for a valid value in the hours preceding the start of the ICU stay (arbitrarily set this amount of time to 32h before the ICU stay). If that didn't fill the gaps at

the beginning of an ICU stay, a backward filling would be applied. Remaining missing values will be filled with a forward fill. In this way all the missing values in laboratory results are filled with the neighbouring values.

2. Filling missing values with a machine learning approach, namely a nearest neighbour imputer. Each sample's missing values are imputed using the mean value from n nearest neighbours found in the training set. Two samples are close if the features that neither is missing are close. By default, a euclidean distance metric that supports missing values is used to find the nearest neighbours.

6.4 Other data transformations

Data is not always in the shape and form that is needed to the AI algorithms. Four kind of major transformations have been applied to the dataset.

1. Data **normalization**. Unification of numerical attributes to similar scales is required in order to eliminate dominating impact of attributes with large numeric ranges on the attributes with small numerical ranges. The simple yet effective min-max scaling has been applied to all vital signs and laboratory measurements to ensure all the values inside a range of 0-1.
2. **One hot-encoding** of categorical variables. Having both categorical and numerical variables in the features set can structurally not fit the type of input that most AI models require. To overcome this issue is common to map the possible values of all the categorical variables to binary features. For example, taking the gender feature that can assume Female or Male values, applying one hot-encoding means to substitute the gender column with Female and Male column and insert a value of 1 under the right column and a 0 in the other.
3. Extracting **meaningful statistical measures** from each time series in order to try to characterize them better. That means that for each vital sign and laboratory result time series the following measures have been calculated: minimum, maximum, median, slope of the regression line. Other studies have found that those measures could characterize the patient while significantly reducing the number of features.
4. Aggregating vital signs and laboratory results into Early Warning Scores. Following the diagram in (Redfern et al., 2018) NEWS score and LDTEWS scores have been calculated and added to the features set.

7 Modelling

7.1 Feature sets

Two steps of data pre-processing (dealing with missing values and extracting meaningful statistical measures) ended up splitting the dataset into two copies. In the end we had four datasets with various characteristics and number of features. The reason why maintaining different copies of the dataset can be useful is simple: different pre-processing approaches lead to different classification results and it is difficult to apriori predict which one will have better performance.

Before looking at the features sets is important to point out that Glucose has been dropped because there's no information about the patient's state at the moment of the measurement (glucose fluctuates too much if the patient is fasted or not and, depending on the patient's condition, could be measured anytime).

Dataset #	Missing Values	Features	# of features
1	Filled	Time series	243
2	Filled	Statistical measures	147
3	Imputed	Time series	243
4	Imputed	Statistical measures	147

Table 7-1 – Different configuration of the dataset by combination of features and missing value's filling strategy

The best feature set in classification resulted to be the number 3, in which missing values have been imputed with a k-Nearest Neighbors imputer and the whole time series have been used for vital signs and laboratory analysis. For this reason, every result that follows will be relative to it.

ICU_ID	Age	Gender	Vital Signs (24x8 values)	Laboratory Analysis (3x7 values)	NEWS	LDT-EWS	Mortality

Figure 7-1 – Final feature set shape

In order to reach the shape in Figure 7-1, some final refinements and reductions have been applied to the dataset:

Data reduction step	# of rows
Full Cohort after pre-processing	4056339
Removing NaN values at the start of the ICUs	4030339
Cutting ICUs at 24h and removing ICUs with more than 25% of missing values	827112
Removing ICUs with no vital sign measurements and/or no laboratory results	30562

Table 7-2 – Last pre-processing steps and total remaining # of rows

Finally, the dataset has been split 80:20 for training and testing.

7.2 Balanced algorithms

Most classification algorithms will only perform optimally when the number of samples of each class is roughly the same. Highly skewed datasets, where the minority is heavily outnumbered by one or more classes (death vs survival), have proven to be a challenge while at the same time becoming more and more common. Ensembling classifiers have shown to improve classification performance compare to single learner. However, they will be affected by class imbalance.

One way of addressing this issue is by re-sampling the dataset as to offset this imbalance with the hope of arriving at a more robust and fair decision boundary than you would otherwise.

Imbalanced-learn is a python package offering a number of re-sampling techniques commonly used in datasets showing strong between-class imbalance. (Lemaître et al., 2017)

The parameters of each of the following AI model implementation have been optimized by the use of a randomized search on hyper-parameters with a 5-fold cross validation. This method achieves similar results to a standard grid search with much less computational effort, hence opening the possibility to try a larger number of parameters combination. In contrast to a classical grid search, not all parameter values are tried out, but rather a fixed number of parameter settings is sampled from the specified distributions. (*Scikit-Learn: Machine Learning in Python — Scikit-Learn 0.23.2 Documentation*, n.d.)

```
from sklearn.model_selection import RandomizedSearchCV

param_grid = {
    'n_estimators': [200, 300, 500, 800, 1000],
    'max_features': ['auto', 'sqrt', 'log2'],
    'criterion' :['gini', 'entropy']
}

brf = BalancedRandomForestClassifier(random_state=42,
                                     n_jobs=-1)

CV_brf = RandomizedSearchCV(estimator=brf, param_distributions=param_grid, cv= 5)
CV_brf.fit(X_train, y_train)
```

Figure 7-2 – Example of randomized search (Balanced Random Forest)

7.2.1 Bagging

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (in our case a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it. (Breiman, 1996)

The BalancedBaggingClassifier model internally uses an additional balancing step to the bagging approach by separately undersampling each bootstrap sample.

Parameter	Value
n_estimators	250

Table 7-3 – Bagging classifier parameters (optimized by randomized search)

7.2.2 Boosting

An AdaBoost (adaptive boosting) (Freund & Schapire, 1997) classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. Easy Ensemble classifier and RUSBoost are natural evolutions of the AdaBoost classifier aimed to better deal with the unbalanced class problem by balancing classes in two different steps of the learning process.

Easy Ensemble (Liu et al., 2009) classifier is an ensemble of AdaBoost learners trained on different balanced bootstrap samples. The balancing is achieved by random under-sampling.

Parameters	Value
Base estimator	AdaBoost
n_estimators (base estimator)	10
n_estimators (Ensemble)	20

Table 7-4 – Easy Ensemble classifier parameters (optimized by randomized search)

In the RUSBoost (Seiffert et al., 2010) classifier, during learning, the problem of class balancing is alleviated by random under-sampling the sample at each iteration of the boosting algorithm.

Parameters	Value
Base estimator	AdaBoost
n_estimators (base estimator)	10
n_estimators (RUSBoost)	50
Learning Rate	0.3

Table 7-5 – RUSBoost classifier parameters (optimized by randomized search)

7.2.3 Balanced Random Forest

In random forests, each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set.

Furthermore, when splitting each node during the construction of a tree, the best split is found either from all input features or a random subset of size max_features.

The purpose of these two sources of randomness is to decrease the variance of the forest estimator. Indeed, individual decision trees typically exhibit high variance and tend to overfit. The injected randomness in forests yield decision trees with somewhat decoupled prediction errors. By taking an average of those predictions, some errors can cancel out. Random forests achieve a reduced variance by combining diverse trees, sometimes at the cost of a slight increase in bias.

In practice the variance reduction is often significant, hence yielding an overall better model. (Chen et al., 2004)

Parameters	Value
n_estimators	800
Max_features	auto
criterion	entropy

Table 7-6 – Balanced Random Forest classifier parameters (optimized by randomized search)

7.3 Models evaluation

Three model evaluation metrics have been used to compare the selected models:

1. Balanced Accuracy: Accuracy means the rate of correct classifications. In the case of a balanced algorithm it is defined as the average of recall (class specific accuracy) obtained on each class. The best value is 1 and the worst value is 0.
2. Geometric Mean: The Geometric Mean (G-Mean) is a metric that measures the balance between classification performances on both the majority and minority classes. A low G-Mean is an indication of a poor performance in the classification of the positive cases even if the negative cases are correctly classified as such.
3. AUROC: It stands for Area Under the Receiver Operator Characteristic. A ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. The AUROC is equal to the probability that a predictor will rank a randomly chosen positive instance higher than a randomly chosen negative one.

The scores have been calculated by averaging scores from a 10-fold cross validation trained on the best hyperparameters.

Classifier	Balanced Accuracy	Geometric Mean	AUROC
Balanced Bagging	0.705	0.698	0.788
Easy Ensemble Classifier	0.711	0.711	0.779
RUSBoost	0.715	0.715	0.787
Balanced Random Forest	0.715	0.715	0.787

Table 7-7 – Classification scores for selected algorithms

As we can see, the performances are pretty similar among the classifiers, with RUSBoost and Balanced Random Forest being the best overall with the same scores. By looking at the confusion matrices of those two algorithms we can see that the Accuracy is really balanced between the two classes. The only point in favour of BRF is that it is much faster than RUSBoost to train.

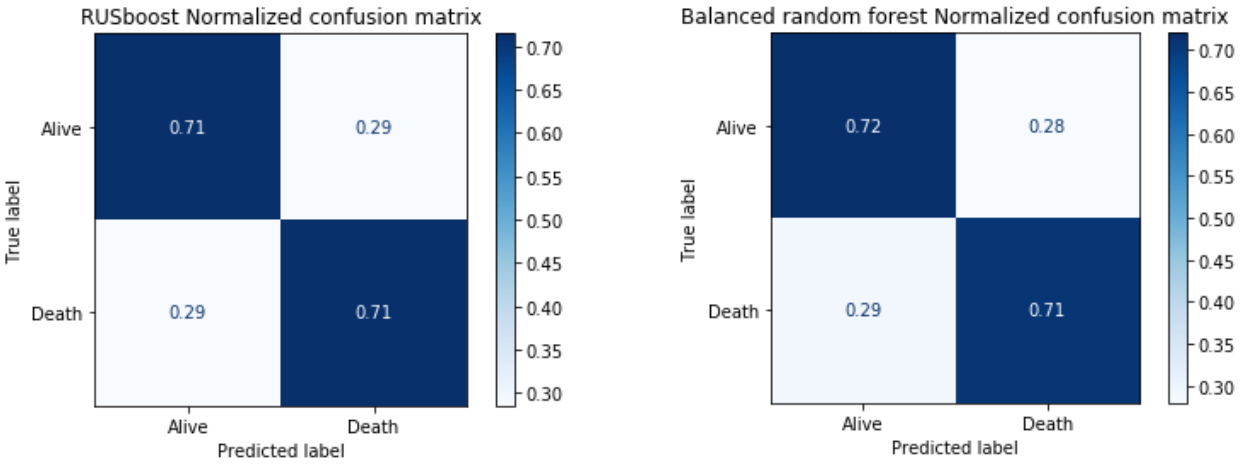


Figure 7-3 – Confusion matrices of the best performing classifiers (RUSBoost on the left - Balanced Random Forest on the right)

8 Conclusions

The more general the outcome to predict, the more difficult it is for a model to generalize the response.

Although a good result was achieved in the prediction using the Balanced Random Forest and RUSBoost algorithms (accuracy of 71.5% and 0.78 AUROC), mortality prediction is far from being a fully automated process. In this context, a failed attempt to classify can spell the death of a patient, which is why such models must be used as support for medical decisions in the ICU and the final word should always be left to the human expert.

Results obtained are slightly inferior to the ones found in the literature (see chapter 3 for some examples): the main reason for that can be found in the complexity of the study and the strict time constraint. Other limiting factors have been a lack of medical background and the impossibility to exploit the University's physical resources due to the Covid-19 pandemic.

Overall, it has been a challenging project that put my skills to the test by improving them and making me deepen my knowledge of DMT and AI.

8.1 Recommendations for future improvements

Below are listed some recommendations for improving and expanding this work:

1. Explore why feature sets that used summary statistical measures instead of the entire time series ended up achieving much worse results
2. Try feature extraction and feature importance techniques
3. Experiment with different sizes of train-test split data
4. Try hybrid undersampling and oversampling techniques before training
5. Try to classify longer time periods (48h – 72h)
6. Implement balanced instances of Deep Learning models (DNN, Convolutional Neural Networks etc.)

References

Aczon, M., Ledbetter, D., Ho, L., Gunny, A., Flynn, A., Williams, J., & Wetzel, R. (2017). *Dynamic Mortality Risk Predictions in Pediatric Critical Care Using Recurrent Neural Networks*. http://arxiv.org/abs/1701.06675

Awad, A., Bader–El–Den, M., & McNicholas, J. (2017). Patient length of stay and mortality prediction: A survey. *Health Services Management Research*. https://doi.org/10.1177/0951484817696212

Barbieri, S., Kemp, J., Perez-Concha, O., Kotwal, S., Gallagher, M., Ritchie, A., & Jorm, L. (2020). Benchmarking Deep Learning Architectures for Predicting Readmission to the ICU and Describing Patients-at-Risk. *Scientific Reports*. https://doi.org/10.1038/s41598-020-58053-z

Breiman, L. (1996). Bagging predictors. *Machine Learning*. https://doi.org/10.1007/bf00058655

Cartin-Ceba, R., Afessa, B., & Gajic, O. (2007). Low baseline serum creatinine concentration predicts mortality in critically ill patients independent of body mass index*. *Critical Care Medicine*, *35*(10), 2420–2423. https://doi.org/10.1097/01.CCM.0000281856.78526.F4

Chen, C., Liaw, A., & Breiman, L. (2004). Using Random Forest to Learn Imbalanced Data | Department of Statistics. In *University of California, Berkeley*.

Churpek, M. M., Adhikari, R., & Edelson, D. P. (2016). The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation*, *102*, 1–5. https://doi.org/10.1016/j.resuscitation.2016.02.005

Churpek, M. M., Yuen, T. C., Winslow, C., Meltzer, D. O., Kattan, M. W., & Edelson, D. P. (2016). Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. *Critical Care Medicine*. https://doi.org/10.1097/CCM.0000000000001571

Citi, L., & Barbieri, R. (2012). PhysioNet 2012 Challenge: Predicting mortality of ICU patients using a cascaded SVM-GLM paradigm. *Computing in Cardiology*.

Clermont, G., Angus, D. C., DiRusso, S. M., Griffin, M., & Linde-Zwirble, W. T. (2001). Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models. *Critical Care Medicine*. https://doi.org/10.1097/00003246-200102000-00012

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*. https://doi.org/10.1023/A:1022627411411

CRISP-DM, still the top methodology for analytics, data mining, or data science projects. (n.d.). Retrieved September 16, 2020, from https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*. https://doi.org/10.1016/j.artmed.2004.07.002

Desautels, T., Calvert, J., Hoffman, J., Jay, M., Kerem, Y., Shieh, L., Shimabukuro, D., Chettipally, U., Feldman, M. D., Barton, C., Wales, D. J., & Das, R. (2016). Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Medical Informatics*. https://doi.org/10.2196/medinform.5909

do Nascimento, G. V. R., Gabriel, D. P., Abrão, J. M. G., & Balbi, A. L. (2010). When is dialysis indicated in acute kidney injury? *Renal Failure*, *32*(3), 396–400. https://doi.org/10.3109/08860221003642633

Doig, G. S., Inman, K. J., Sibbald, W. J., Martin, C. M., & Robertson, J. M. (1993). Modeling

mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression. *Proceedings / the ... Annual Symposium on Computer Application [Sic] in Medical Care. Symposium on Computer Applications in Medical Care*.

Dybowski, R., Weller, P., Chang, R., & Gant, V. (1996). Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. *Lancet*. https://doi.org/10.1016/S0140-6736(96)90609-1

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*. https://doi.org/10.1006/jcss.1997.1504

Gaspar, J., Catumbela, E., Marques, B., & Freitas, A. (2011). A systematic review of outliers detection techniques in medical data: Preliminary study. *HEALTHINF 2011 - Proceedings of the International Conference on Health Informatics*.

Ge, W., Huh, J. W., Park, Y. R., Lee, J. H., Kim, Y. H., & Turchin, A. (2018). An Interpretable ICU Mortality Prediction Model Based on Logistic Regression and Recurrent Neural Networks with LSTM units. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*.

Giudici, P., & Figini, S. (2009). Applied Data Mining for Business and Industry. In *Applied Data Mining for Business and Industry*. https://doi.org/10.1002/9780470745830

Goldman, A. I. (2015). Theory of Human Action. In *Theory of Human Action*. https://doi.org/10.1515/9781400868971

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and TechniquesHan, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. San Francisco, CA, itd: Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00001-0. In *San Francisco, CA, itd: Morgan Kaufmann*. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., & Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. *Scientific Data*. https://doi.org/10.1038/s41597-019-0103-9

Henriques, J. H., & Rocha, T. R. (2009). Prediction of acute hypotensive episodes using neural network multi-models. *Computers in Cardiology*.

Higgins, T. L. (2007). Quantifying risk and benchmarking performance in the adult intensive care unit. In *Journal of Intensive Care Medicine*. https://doi.org/10.1177/0885066607299520

Hofer, I. S., Lee, C., Gabel, E., Baldi, P., & Cannesson, M. (2020). Development and validation of a deep neural network model to predict postoperative mortality, acute kidney injury, and reintubation using a single feature set. *Npj Digital Medicine*. https://doi.org/10.1038/s41746-020-0248-0

Jarvis, S. W., Kovacs, C., Briggs, J., Meredith, P., Schmidt, P. E., Featherstone, P. I., Prytherch, D. R., & Smith, G. B. (2015). Are observation selection methods important when comparing early warning score performance? *Resuscitation*. https://doi.org/10.1016/j.resuscitation.2015.01.033

Jensen, M. T., Suadicani, P., Hein, H. O., & Gyntelberg, F. (2013). Elevated resting heart rate, physical fitness and all-cause mortality: A 16-year follow-up in the Copenhagen Male Study. *Heart*. https://doi.org/10.1136/heartjnl-2012-303375

Jo, Y., Lee, L., & Palaskar, S. (2017). *Combining LSTM and Latent Topic Modeling for Mortality Prediction*. http://arxiv.org/abs/1709.02842

Johnson, A. E. W., Stone, D. J., Celi, L. A., & Pollard, T. J. (2018). The MIMIC Code Repository: Enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, *25*(1), 32–39. https://doi.org/10.1093/jamia/ocx084

Jung, S. M., Kim, Y. J., Ryoo, S. M., & Kim, W. Y. (2019). Relationship between low hemoglobin levels and mortality in patients with septic shock. *Acute and Critical Care*, *34*(2), 141–147. https://doi.org/10.4266/acc.2019.00465

Kause, J., Smith, G., Prytherch, D., Parr, M., Flabouris, A., & Hillman, K. (2004). A comparison of Antecedents to Cardiac Arrests, Deaths and EMergency Intensive care Admissions in Australia and New Zealand, and the United Kingdom - The ACADEMIA study. *Resuscitation*, *62*(3), 275–282. https://doi.org/10.1016/j.resuscitation.2004.05.016

Kellett, J. (2017). The Assessment and Interpretation of Vital Signs. In *Textbook of Rapid Response Systems*. https://doi.org/10.1007/978-3-319-39391-9_8

Kellett, J., & Sebat, F. (2017). Make vital signs great again – A call for action. *European Journal of Internal Medicine*. https://doi.org/10.1016/j.ejim.2017.09.018

Kim, S., Kim, W., & Woong Park, R. (2011). A comparison of intensive care unit mortality prediction models through the use of data mining techniques. *Healthcare Informatics Research*. https://doi.org/10.4258/hir.2011.17.4.232

Knaus, W. A., Draper, E. A., Wagner, D. P., & Zimmerman, J. E. (1985). APACHE II: A severity of disease classification system. *Critical Care Medicine*. https://doi.org/10.1097/00003246-198510000-00009

Larose, D. T., & Larose, C. D. (2014). Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition. In *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition*. https://doi.org/10.1002/9781118874059

Le Gall, J. R. (1993). A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA: The Journal of the American Medical Association*. https://doi.org/10.1001/jama.270.24.2957

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*.

Lemeshow, S., Gehlbach, S. H., Klar, J., Avrunin, J. S., Teres, D., & Rapoport, J. (1993). Mortality Probability Models (MPM II) Based on an International Cohort of Intensive Care Unit Patients. *JAMA: The Journal of the American Medical Association*. https://doi.org/10.1001/jama.1993.03510200084037

Lewicki, P., & Hill, T. (2006). Statistics : Methods and Applications - A comprehensive reference for science, industry and data mining. In *StatSoft Inc.* https://doi.org/10.1016/B978-0-323-03707-5.50024-3

Liu, X. Y., Wu, J., & Zhou, Z. H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*. https://doi.org/10.1109/TSMCB.2008.2007853

Maheshwari, K., Nathanson, B. H., Munson, S. H., Khangulov, V., Stevens, M., Badani, H., Khanna, A. K., & Sessler, D. I. (2018). The relationship between ICU hypotension and in-hospital mortality and morbidity in septic patients. *Intensive Care Medicine*, *44*(6), 857–867. https://doi.org/10.1007/s00134-018-5218-5

Maslove, D. M., Dubin, J. A., Shrivats, A., & Lee, J. (2016). Errors, Omissions, and Outliers in Hourly Vital Signs Measurements in Intensive Care. *Critical Care Medicine*. https://doi.org/10.1097/CCM.0000000000001862

Meyfroidt, G., Güiza, F., Ramon, J., & Bruynooghe, M. (2009). Machine learning techniques to examine large patient databases. In *Best Practice and Research: Clinical Anaesthesiology*. https://doi.org/10.1016/j.bpa.2008.09.003

MIMIC Critical Care Database. (n.d.). Retrieved September 16, 2020, from

https://mimic.physionet.org/

Montgomery, D. C. (2013). Design and Analysis of Experiments (International Student Version). In *John Wiley & Sons, Inc*. https://doi.org/10.1198/tech.2006.s372

Morgan, R., Williams, F., & Wright, M. (1997). An early warning scoring system for detecting developing critical illness. In *Clin Intensive Care*.

Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., & Buchman, T. G. (2018). An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Critical Care Medicine*. https://doi.org/10.1097/CCM.0000000000002936

Palmer, A., Jimenez, R., & Gervill, E. (2011). Data Mining: Machine Learning and Statistical Techniques. In *Knowledge-Oriented Applications in Data Mining*. https://doi.org/10.5772/13621

Peterson, L. (2009). K-nearest neighbor. *Scholarpedia*. https://doi.org/10.4249/scholarpedia.1883

Pirracchio, R., Petersen, M. L., Carone, M., Rigon, M. R., Chevret, S., & van der Laan, M. J. (2015). Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): A population-based study. *The Lancet Respiratory Medicine*. https://doi.org/10.1016/S2213-2600(14)70239-5

Prytherch, D. R., Smith, G. B., Schmidt, P. E., & Featherstone, P. I. (2010). ViEWS-Towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation*. https://doi.org/10.1016/j.resuscitation.2010.04.014

Ramon, J., Fierens, D., Güiza, F., Meyfroidt, G., Blockeel, H., Bruynooghe, M., & Van Den Berghe, G. (2007). Mining data from intensive care patients. *Advanced Engineering Informatics*, *21*(3), 243–256. https://doi.org/10.1016/j.aei.2006.12.002

Ranzani, O. T., Zampieri, F. G., Forte, D. N., Azevedo, L. C. P., & Park, M. (2013). C-Reactive Protein/Albumin Ratio Predicts 90-Day Mortality of Septic Patients. *PLoS ONE*, *8*(3), e59321. https://doi.org/10.1371/journal.pone.0059321

Redfern, O. C., Pimentel, M. A. F., Prytherch, D., Meredith, P., Clifton, D. A., Tarassenko, L., Smith, G. B., & Watkinson, P. J. (2018). Predicting in-hospital mortality and unanticipated admissions to the intensive care unit using routinely collected blood tests and vital signs: Development and validation of a multivariable model. *Resuscitation*. https://doi.org/10.1016/j.resuscitation.2018.09.021

Ribas, V. J., Lopez, J. C., Ruiz-Sanmartin, A., Ruiz-Rodriguez, J. C., Rello, J., Wojdel, A., & Vellido, A. (2011). Severe sepsis mortality prediction with relevance vector machines. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. https://doi.org/10.1109/IEMBS.2011.6089906

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*. https://doi.org/10.1038/323533a0

Sadeghi, R., Banerjee, T., & Romine, W. (2018). Early hospital mortality prediction using vital signals. *Smart Health*. https://doi.org/10.1016/j.smhl.2018.07.001

scikit-learn: machine learning in Python — scikit-learn 0.23.2 documentation. (n.d.). Retrieved September 21, 2020, from https://scikit-learn.org/stable/

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*. https://doi.org/10.1109/TSMCA.2009.2029559

Silva, Á., Cortez, P., Santos, M. F., Gomes, L., & Neves, J. (2006). Mortality assessment in intensive care units via adverse events using artificial neural networks. *Artificial Intelligence in Medicine*. https://doi.org/10.1016/j.artmed.2005.07.006

Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E., & Featherstone, P. I. (2013). The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*. https://doi.org/10.1016/j.resuscitation.2012.12.016

Tongyoo, S., Viarasilpa, T., & Permpikul, C. (2018). Serum potassium levels and outcomes in critically ill patients in the medical intensive care unit. *Journal of International Medical Research*, *46*(3), 1254–1262. https://doi.org/10.1177/0300060517744427

Vincent, J. L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C. K., Suter, P. M., & Thijs, L. G. (1996). The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*. https://doi.org/10.1007/BF01709751

Waheed, U., Williams, P., Brett, S., Baldock, G., & Soni, N. (2003). White cell count and intensive care unit outcome. *Anaesthesia*, *58*(2), 180–182. https://doi.org/10.1046/j.1365-2044.2003.02964_5.x

Wong, L. S. S., & Young, J. D. (1999). A comparison of ICU mortality prediction using the APACHE II scoring system and artificial neural networks. *Anaesthesia*. https://doi.org/10.1046/j.1365-2044.1999.01104.x

Zhou, D., Li, Z., Shi, G., & Zhou, J. (2020). Time spent in oxygen saturation 95-99% is associated with reduced mortality in critically ill patients with mechanical ventilation. In *Critical Care* (Vol. 24, Issue 1, p. 414). BioMed Central. https://doi.org/10.1186/s13054-020-03126-8

Appendix A1: Project Initiation Document

School of Computing
Postgraduate Programme

MSc in Information Systems

Project Specification

Davide Garofalo

Project Specification

1. Basic details

Student name:	Davide Garofalo
Draft project title:	Predicting mortality in ICU with artificial intelligence
Course and year:	MSc in Information Systems – PJE60 – 2019/2020
Client organisation:	University of Portsmouth
Client contact name:	Professor Jim Briggs
Project supervisor:	Professor Jim Briggs

2. Outline of the project environment

Who is the client? What do they do?

The client is Professor Jim Briggs. He is Professor of Informatics in the University of Portsmouth and Director of the Centre for Healthcare Modelling and Informatics (CHMI). Professor Briggs, along with Professor David Prytherch and their research team, are best known for their contributions to clinical outcome modelling in the area of patient deterioration.

What is their problem? Why does it need to be solved?

Predicting mortality in ICU is something that, if improved, not only can save more lives by escalating targeted care for who needs it but can also save big sums of money for public healthcare by optimizing hospital resources and nurse workload.

3. The problem to be solved

Give more detail about the problem.

Despite being one of the most common clinical practices, vital sign monitoring is still a field open for improvement on multiple sides. In recent times, the adoption of early warning scores has been highly effective for detecting patients at risk of clinical deterioration or death, prompting a timelier clinical response, with the aim of improving patient outcomes in the NHS.
With the current study project, I'm going to explore how the adoption of cutting-edge technologies and AI algorithms could improve the outcome of early mortality prediction, especially in ICU.
If good results are reached in a timely manner, the big issue of standardizing the frequency of observation (FOBS) with which medical staff should take measurement of vital signs will be tackled.

What are the aims and objectives of the project?

The aim of this project is to build an artificial intelligence model that has good accuracy in predicting ICU mortality. The objectives that must be achieved to reach this final goal are:

- Explore the MIMIC-III database and getting useful insights
- Through intense literature research, acquire basic knowledge of vital signs monitoring and existing artificial intelligence techniques used in this field
- Extract a meaningful cohort that could be representative for the general population
- Apply machine learning and deep learning techniques to predict mortality in ICU
- Investigate various techniques for imputing missing data in order to have similar prediction with fewer time points (if time permits)

What constraints are there on your solution to the problem?

- No previous background medical knowledge
- Limited time
- Ambitious and complex tasks

4. Breakdown of tasks

What is going to be your approach?

To optimize the most important and scarce resource of the project (Time) I'm going to follow the CRISP-DM methodology (Wirth & Hipp, 2000) and a strict and accurate planning schedule.

What background research do you need to do?

Background research for this project can be split in two blocks:

1. Healthcare related knowledge, regarding vital sign monitoring, its values and practices, the usage of aggregated early warning scores etc.
2. Artificial intelligence and data mining related knowledge, regarding the best models used in the last years for similar classification problems and best practices to handle health related data.

What tools do you need to acquire?

No particular tools are needed for the project, except for a laptop with the required software and a local implementation of the MIMIC-III database. The software includes R Studio, Python, Jupyter Notebook and PostgreSQL.

What skills do you require and how are you going to acquire those that you do not already have?

Skills and knowledge already possessed:

- SQL querying
- Python data pre-processing (using libraries such as NumPy, pandas, matplotlib)

- R for data visualization and exploration
- Basic statistical analysis

Skills and knowledge that have to be further improved by researching and practicing:

- Features selection and extraction techniques
- Resampling and other class balancing techniques
- Artificial intelligence algorithms for binary and probabilistic classification

What do you need to design and build?

Knowledge of the domain and the data (MIMIC-III data schema and description of tables)

A laptop with all the required software (R Studio, Python, Jupyter Notebook, PostgreSQL)

Due to the size of the data to be analysed and the required high computational power, free cloud-based services will be used:

1. A cloud-based querying framework with an instance of the MIMIC-III database for the most expensive queries (Google Big Query)
2. A cloud-based computing platform such as Google Colab for training the AI models

5. Project deliverables

What information system artefacts will be developed/adapted? What documents will be produced?

- Cohort data in csv format will be extracted from the database as result of the data cleaning and handling process.
- Infographics and reports will be documented as result of the exploratory data analysis step.
- Jupyter Notebooks to train, assess and save AI models will be made in order to share coding and classification results with supervisors.
- A final dissertation document will be produced to share all the project steps, findings and results.

6. Requirements

What are the client's requirements?

To get reasonable prediction rate (65-70% accuracy) or, in the case of failure, bring strong evidence and motivation for the poor result (e.g. insufficient data, insufficient time to build/understand complex systems etc.) and viable suggestions for improvement.

Legal, ethical, professional, social issues
What are the legal/ethical/professional/social issues that may impose constraints on the project? How will you ensure that they will be complied with, or what steps will you take to avoid/mitigate their effects?

In US and UK, legal frameworks such as 'HIPAA Privacy and Security rules' [HIPAA] and General Data Protection Regulation [GDPR] are implemented to ensure data subjects' protection rights. The MIMIC-III database is HIPAA compliant and to use it is mandatory to complete an MIT course on human research and data or specimen handling.
The researcher is also asked to exercise all reasonable and prudent care to avoid disclosure of identities and maintain the physical and electronic security of the data.

What research ethics approval (if any) is needed for your work and what will you do to obtain it?

Citi certificate of approval is attached

7. Facilities and resources

What computing/IT facilities will you use/require? What other facilities/resources will you use/require?

The project doesn't require external or additional resources other than personal laptop, free cloud computing platforms (BigQuery and Google Colab), research literature (provided by supervisor and university library), supervisors' knowledge and feedbacks.

8. Project plan

What are you going to do when?

Weekly schedule is attached

What risks to the success of the project have you identified? What steps can you take to minimise them? What backup plans do you have if identified things go wrong?

Risk	Probability (Low-Medium-High)	Impact (Low-Medium-High)	Prevention
Loss of work	Low	High	Keep continuous backups regularly. Save the work locally and on the cloud.
Time shortage	High	High	Have a clear and detailed schedule to ensure optimal time management
Suboptimal Results	Medium	Low	Trace and motivate every result, especially the negative ones, to understand if they

			arise from intrinsic project limitations or resource limitations
Technical Issues	High	Medium	Pre-allocate days for handling the ever-present technical issues that arise from coding work

9. Project mode

If there are two possibilities for your project mode, after negotiation, please record your planned duration and submission date. It is also helpful to record your initial registration mode (i.e. are you a full time or a part time student). Remember, the exact dates will be announced through Moodle – these represent a generic guideline.

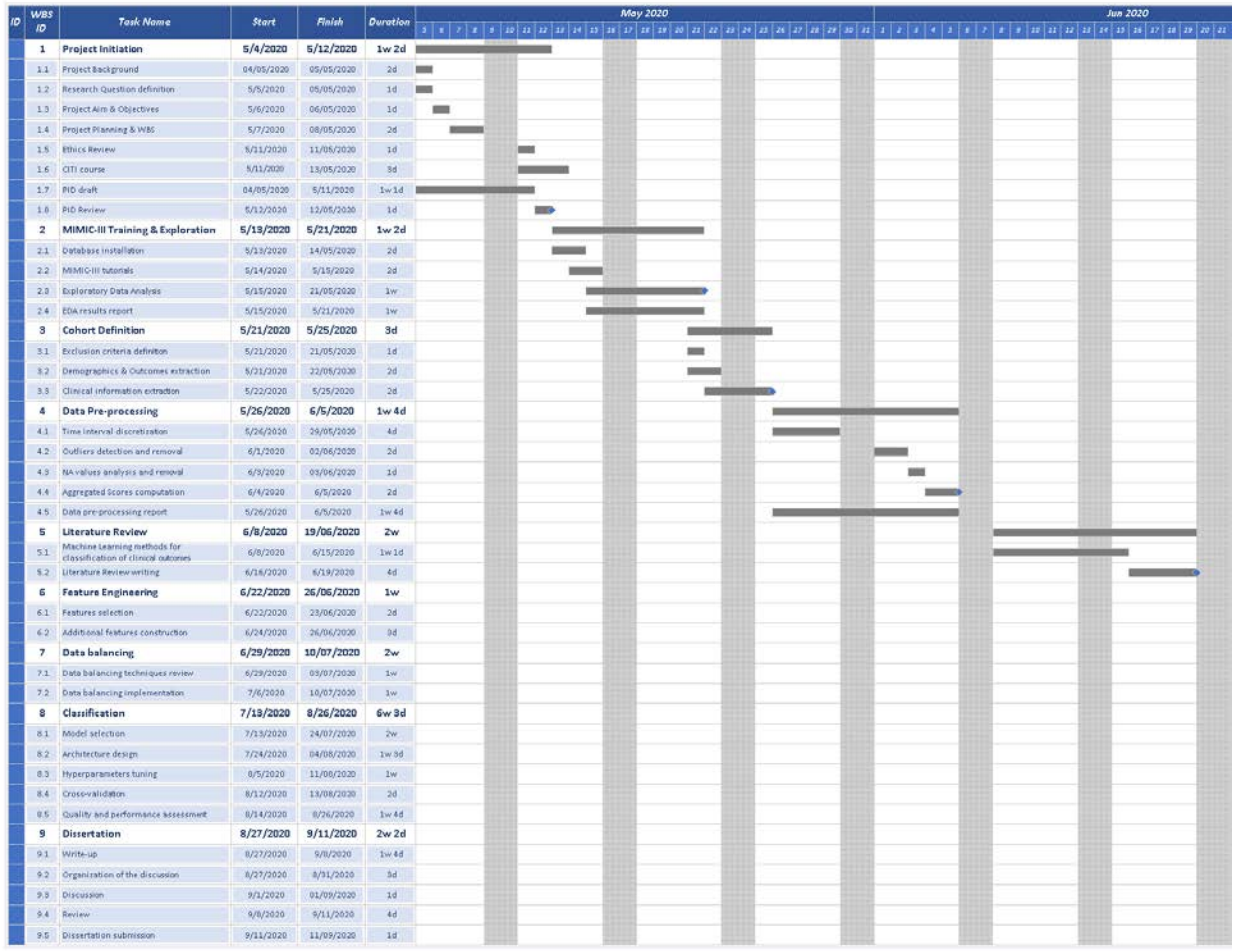
	Please delete as appropriate	
Registration mode	Full Time	
Project mode	Full Time	
Planned submission deadline	01/09/20	

10. Signatures

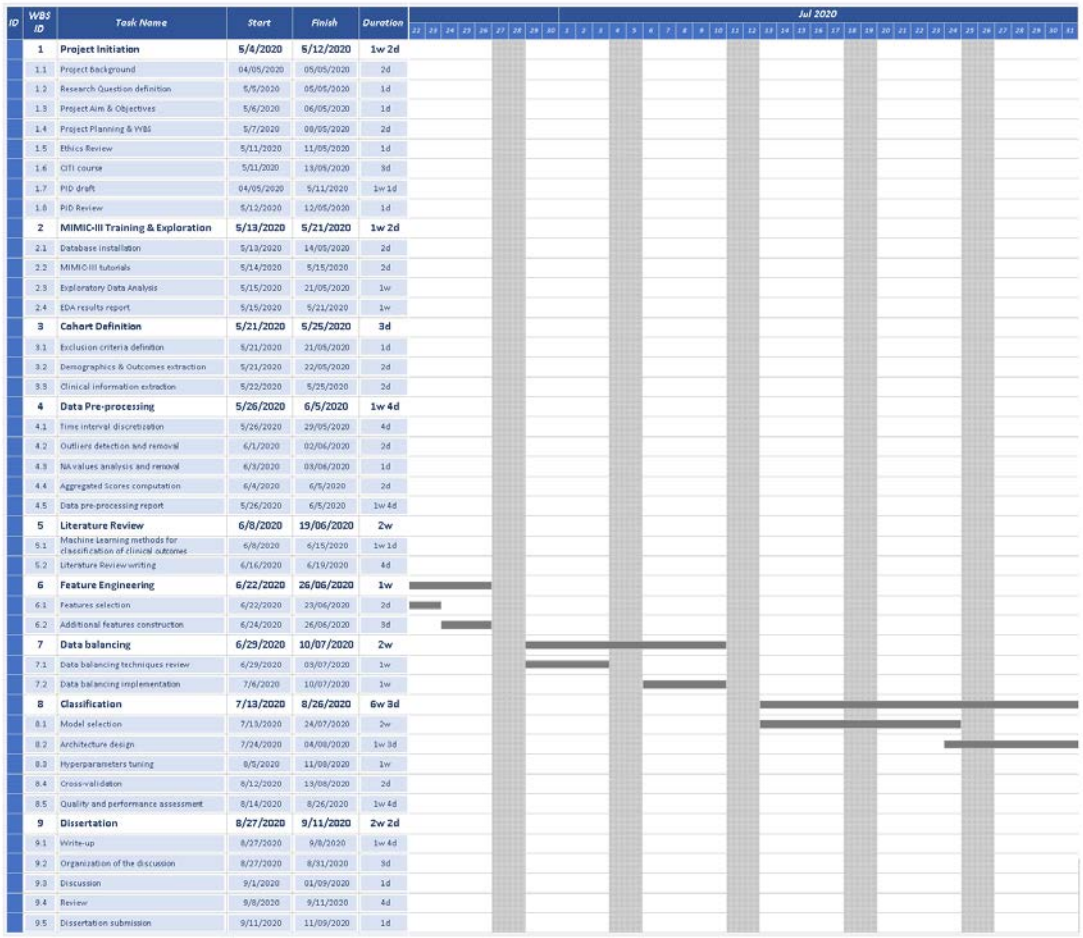
	Signature:	Date:
Student		15/06/2020
Client	J S Briggs	18/06/2020
Project supervisor	J S Briggs	18/06/2020

Confirmed by e-mail on the 18/06/2020

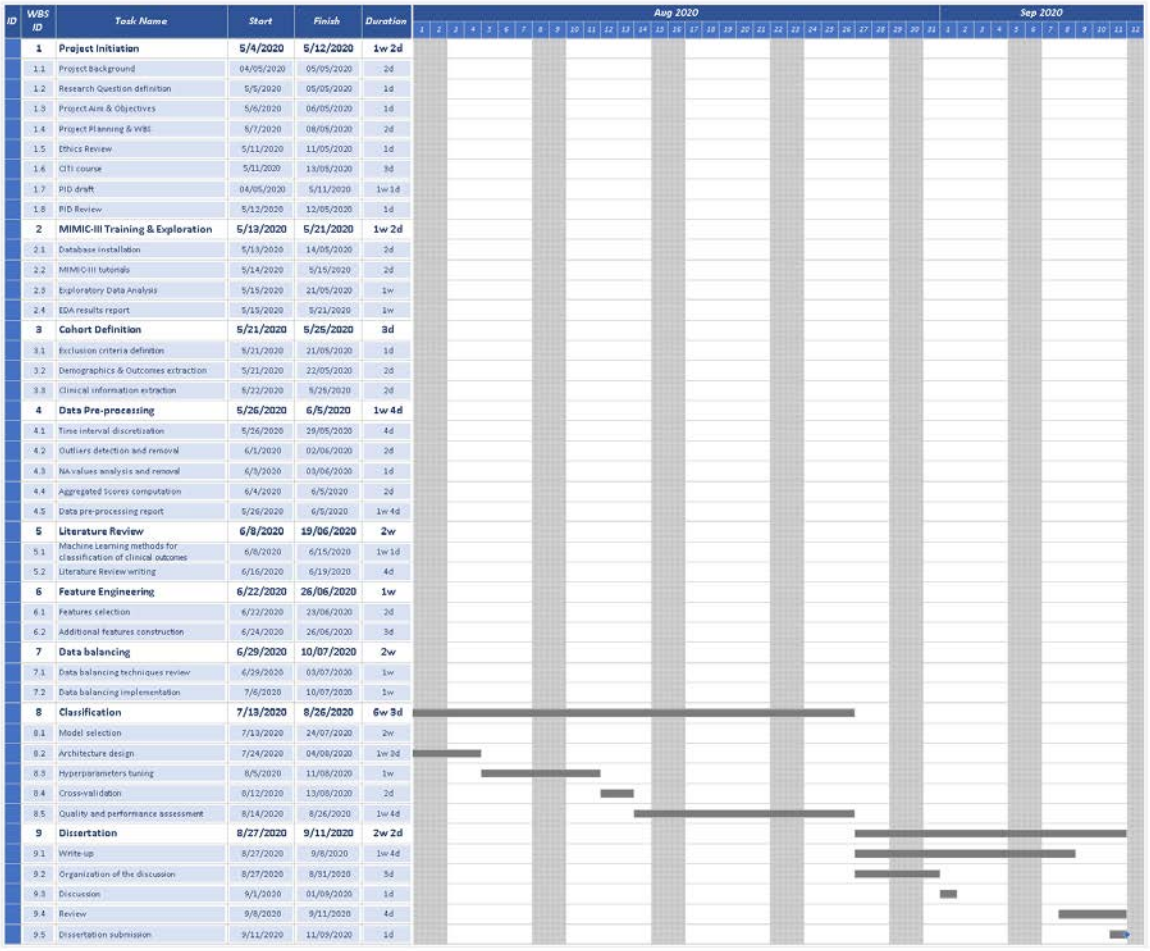
Appendix A2: Gantt Diagram



ID	WBS ID	Task Name	Start	Finish	Duration
	1	**Project Initiation**	**5/4/2020**	**5/12/2020**	**1w 2d**
	1.1	Project Background	04/05/2020	05/05/2020	2d
	1.2	Research Question definition	5/5/2020	05/05/2020	1d
	1.3	Project Aim & Objectives	5/6/2020	06/05/2020	1d
	1.4	Project Planning & WBS	5/7/2020	08/05/2020	2d
	1.5	Ethics Review	5/11/2020	11/05/2020	1d
	1.6	CITI course	5/11/2020	13/05/2020	3d
	1.7	PID draft	04/05/2020	5/11/2020	1w 1d
	1.8	PID Review	5/12/2020	12/05/2020	1d
	2	**MIMIC-III Training & Exploration**	**5/13/2020**	**5/21/2020**	**1w 2d**
	2.1	Database installation	5/13/2020	14/05/2020	2d
	2.2	MIMIC-III tutorials	5/14/2020	5/15/2020	2d
	2.3	Exploratory Data Analysis	5/15/2020	21/05/2020	1w
	2.4	EDA results report	5/15/2020	5/21/2020	1w
	3	**Cohort Definition**	**5/21/2020**	**5/25/2020**	**3d**
	3.1	Exclusion criteria definiton	5/21/2020	21/05/2020	1d
	3.2	Demographics & Outcomes extraction	5/21/2020	22/05/2020	2d
	3.3	Clinical information extraction	5/22/2020	5/25/2020	2d
	4	**Data Pre-processing**	**5/26/2020**	**6/5/2020**	**1w 4d**
	4.1	Time interval discretization	5/26/2020	29/05/2020	4d
	4.2	Outliers detection and removal	6/1/2020	02/06/2020	2d
	4.3	NA values analysis and removal	6/3/2020	03/06/2020	1d
	4.4	Aggregated Scores computation	6/4/2020	6/5/2020	2d
	4.5	Data pre-processing report	5/26/2020	6/5/2020	1w 4d
	5	**Literature Review**	**6/8/2020**	**19/06/2020**	**2w**
	5.1	Machine Learning methods for classification of clinical outcomes	6/8/2020	6/15/2020	1w 1d
	5.2	Literature Review writing	6/16/2020	6/19/2020	4d
	6	**Feature Engineering**	**6/22/2020**	**26/06/2020**	**1w**
	6.1	Features selection	6/22/2020	23/06/2020	2d
	6.2	Additional features construction	6/24/2020	26/06/2020	3d
	7	**Data balancing**	**6/29/2020**	**10/07/2020**	**2w**
	7.1	Data balancing techniques review	6/29/2020	03/07/2020	1w
	7.2	Data balancing implementation	7/6/2020	10/07/2020	1w
	8	**Classification**	**7/13/2020**	**8/26/2020**	**6w 3d**
	8.1	Model selection	7/13/2020	24/07/2020	2w
	8.2	Architecture design	7/24/2020	04/08/2020	1w 3d
	8.3	Hyperparameters tuning	8/5/2020	11/08/2020	1w
	8.4	Cross-validation	8/12/2020	13/08/2020	2d
	8.5	Quality and performance assessment	8/14/2020	8/26/2020	1w 4d
	9	**Dissertation**	**8/27/2020**	**9/11/2020**	**2w 2d**
	9.1	Write-up	8/27/2020	9/8/2020	1w 4d
	9.2	Organization of the discussion	8/27/2020	8/31/2020	3d
	9.3	Discussion	9/1/2020	01/09/2020	1d
	9.4	Review	9/8/2020	9/11/2020	4d
	9.5	Dissertation submission	9/11/2020	11/09/2020	1d



ID	WBS ID	Task Name	Start	Finish	Duration
	1	**Project Initiation**	5/4/2020	5/12/2020	1w 2d
	1.1	Project Background	04/05/2020	05/05/2020	2d
	1.2	Research Question definition	5/5/2020	05/05/2020	1d
	1.3	Project Aim & Objectives	5/6/2020	06/05/2020	1d
	1.4	Project Planning & WBS	5/7/2020	08/05/2020	2d
	1.5	Ethics Review	5/11/2020	11/05/2020	1d
	1.6	CITI course	5/11/2020	13/05/2020	3d
	1.7	PID draft	04/05/2020	5/11/2020	1w 1d
	1.8	PID Review	5/12/2020	12/05/2020	1d
	2	**MIMIC-III Training & Exploration**	5/13/2020	5/21/2020	1w 2d
	2.1	Database installation	5/13/2020	14/05/2020	2d
	2.2	MIMIC-III tutorials	5/14/2020	5/15/2020	2d
	2.3	Exploratory Data Analysis	5/15/2020	21/05/2020	1w
	2.4	EDA results report	5/15/2020	5/21/2020	1w
	3	**Cohort Definition**	5/21/2020	5/25/2020	3d
	3.1	Exclusion criteria definition	5/21/2020	21/05/2020	1d
	3.2	Demographics & Outcomes extraction	5/21/2020	22/05/2020	2d
	3.3	Clinical information extraction	5/22/2020	5/25/2020	2d
	4	**Data Pre-processing**	5/26/2020	6/5/2020	1w 4d
	4.1	Time interval discretization	5/26/2020	29/05/2020	4d
	4.2	Outliers detection and removal	6/1/2020	02/06/2020	2d
	4.3	NA values analysis and removal	6/3/2020	03/06/2020	1d
	4.4	Aggregated Scores computation	6/4/2020	6/5/2020	2d
	4.5	Data pre-processing report	5/26/2020	6/5/2020	1w 4d
	5	**Literature Review**	6/8/2020	19/06/2020	2w
	5.1	Machine Learning methods for classification of clinical outcomes	6/8/2020	6/15/2020	1w 1d
	5.2	Literature Review writing	6/16/2020	6/19/2020	4d
	6	**Feature Engineering**	6/22/2020	26/06/2020	1w
	6.1	Features selection	6/22/2020	23/06/2020	2d
	6.2	Additional features construction	6/24/2020	26/06/2020	3d
	7	**Data balancing**	6/29/2020	10/07/2020	2w
	7.1	Data balancing techniques review	6/29/2020	03/07/2020	1w
	7.2	Data balancing implementation	7/6/2020	10/07/2020	1w
	8	**Classification**	7/13/2020	8/26/2020	6w 3d
	8.1	Model selection	7/13/2020	24/07/2020	2w
	8.2	Architecture design	7/24/2020	04/08/2020	1w 3d
	8.3	Hyperparameters tuning	8/5/2020	11/08/2020	1w
	8.4	Cross-validation	8/12/2020	13/08/2020	2d
	8.5	Quality and performance assessment	8/14/2020	8/26/2020	1w 4d
	9	**Dissertation**	8/27/2020	9/11/2020	2w 2d
	9.1	Write-up	8/27/2020	9/8/2020	1w 4d
	9.2	Organization of the discussion	8/27/2020	8/31/2020	3d
	9.3	Discussion	9/1/2020	01/09/2020	1d
	9.4	Review	9/8/2020	9/11/2020	4d
	9.5	Dissertation submission	9/11/2020	11/09/2020	1d



ID	WBS ID	Task Name	Start	Finish	Duration
	1	**Project Initiation**	**5/4/2020**	**5/12/2020**	**1w 2d**
	1.1	Project Background	04/05/2020	05/05/2020	2d
	1.2	Research Question definition	5/5/2020	05/05/2020	1d
	1.3	Project Aim & Objectives	5/6/2020	06/05/2020	1d
	1.4	Project Planning & WBS	5/7/2020	08/05/2020	2d
	1.5	Ethics Review	5/11/2020	11/05/2020	1d
	1.6	CITI course	5/11/2020	13/05/2020	3d
	1.7	PID draft	04/05/2020	5/11/2020	1w 1d
	1.8	PID Review	5/12/2020	12/05/2020	1d
	2	**MIMIC-III Training & Exploration**	**5/13/2020**	**5/21/2020**	**1w 2d**
	2.1	Database installation	5/13/2020	14/05/2020	2d
	2.2	MIMIC-III tutorials	5/14/2020	5/15/2020	2d
	2.3	Exploratory Data Analysis	5/15/2020	21/05/2020	1w
	2.4	EDA results report	5/15/2020	5/21/2020	1w
	3	**Cohort Definition**	**5/21/2020**	**5/25/2020**	**3d**
	3.1	Exclusion criteria definition	5/21/2020	21/05/2020	1d
	3.2	Demographics & Outcomes extraction	5/21/2020	22/05/2020	2d
	3.3	Clinical information extraction	5/22/2020	5/25/2020	2d
	4	**Data Pre-processing**	**5/26/2020**	**6/5/2020**	**1w 4d**
	4.1	Time interval discretization	5/26/2020	29/05/2020	4d
	4.2	Outliers detection and removal	6/1/2020	02/06/2020	2d
	4.3	NA values analysis and removal	6/3/2020	03/06/2020	1d
	4.4	Aggregated Scores computation	6/4/2020	6/5/2020	2d
	4.5	Data pre-processing report	5/26/2020	6/5/2020	1w 4d
	5	**Literature Review**	**6/8/2020**	**19/06/2020**	**2w**
	5.1	Machine Learning methods for classification of clinical outcomes	6/8/2020	6/15/2020	1w 1d
	5.2	Literature Review writing	6/16/2020	6/19/2020	4d
	6	**Feature Engineering**	**6/22/2020**	**26/06/2020**	**1w**
	6.1	Features selection	6/22/2020	23/06/2020	2d
	6.2	Additional features construction	6/24/2020	26/06/2020	3d
	7	**Data balancing**	**6/29/2020**	**10/07/2020**	**2w**
	7.1	Data balancing techniques review	6/29/2020	03/07/2020	1w
	7.2	Data balancing implementation	7/6/2020	10/07/2020	1w
	8	**Classification**	**7/13/2020**	**8/26/2020**	**6w 3d**
	8.1	Model selection	7/13/2020	24/07/2020	2w
	8.2	Architecture design	7/24/2020	04/08/2020	1w 3d
	8.3	Hyperparameters tuning	8/5/2020	11/08/2020	1w
	8.4	Cross-validation	8/12/2020	13/08/2020	2d
	8.5	Quality and performance assessment	8/14/2020	8/26/2020	1w 4d
	9	**Dissertation**	**8/27/2020**	**9/11/2020**	**2w 2d**
	9.1	Write-up	8/27/2020	9/8/2020	1w 4d
	9.2	Organization of the discussion	8/27/2020	8/31/2020	5d
	9.3	Discussion	9/1/2020	01/09/2020	1d
	9.4	Review	9/8/2020	9/11/2020	4d
	9.5	Dissertation submission	9/11/2020	11/09/2020	1d

Appendix B: Certificate of Ethics review

Certificate of Ethics Review

Project Title: Predicting mortality in ICU with artificial intelligence

Name: Davide Garofalo **User ID:** 963994 **Application Date:** 02-Jun-2020 10:38 **ER Number:** ETHIC-2020-686

You must download your referral certificate, print a copy and keep it as a record of this review.

The FEC representative for the School of Computing is Carl Adams

It is your responsibility to follow the University Code of Practice on Ethical Standards and any Department/School or professional guidelines in the conduct of your study including relevant guidelines regarding health and safety of researchers including the following:

- University Policy
- Safety on Geological Fieldwork

All projects involving human participants need to offer sufficient information to potential participants to enable them to make a decision. Template participant information sheets are available from the:

- Univeristy's Ethics Site (Participant information template).

It is also your responsibility to follow University guidance on Data Protection Policy:

- General guidance for all data protection issues
- University Data Protection Policy

Which school/department do you belong to?: **SOC**
What is your primary role at the University?: **PostgraduateStudent**
What is the name of the member of staff who is responsible for supervising your project?: **Jim Briggs**
Is the study likely to involve human subjects (observation) or participants?: **Yes**
Will peoples` involvement be limited to just responding to questionnaires or surveys, or providing structured feedback during software prototyping?: **Yes**
Confirm whether and explain how you will use participant information sheets and apply informed consent.: **Data for this project comes from the MIMIC-III database. The MIMIC project was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). The requirement for individual patient consent was waived because the project did not impact clinical care and all protected health information was deidentified.**
Confirm whether and explain how you will maintain participant anonymity and confidentiality of data collected.: **Before data was incorporated into the MIMIC-III database, it was first de-identified in accordance with Health Insurance Portability and Accountability Act (HIPAA) standards using structured data cleansing and date shifting. The de-identification process for structured data required the removal of all eighteen of the identifying data elements listed in HIPAA.**
Will the study involve National Health Service patients or staff?: **No**
Do human participants/subjects take part in studies without their knowledge/consent at the time, or will deception of any sort be involved? (e.g. covert observation of people, especially if in a non-public place): **No**
Will you collect or analyse personally identifiable information about anyone or monitor their communications or on-line activities without their explicit consent?: **No**
Does the study involve participants who are unable to give informed consent or in are in a dependent position (e.g. children, people with learning disabilities, unconscious patients, Portsmouth University students)?: **No**
Are drugs, placebos or other substances (e.g. food substances, vitamins) to be administered to the study participants?: **No**
Will blood or tissue samples be obtained from participants?: **No**
Is pain or more than mild discomfort likely to result from the study?: **No**
Could the study induce psychological stress or anxiety in participants or third parties?: **No**
Will the study involve prolonged or repetitive testing?: **No**
Will financial inducements (other than reasonable expenses and compensation for time) be offered to participants?: **No**
Are there risks of significant damage to physical and/or ecological environmental features?: **No**
Are there risks of significant damage to features of historical or cultural heritage (e.g. impacts of study techniques, taking of samples)?: **No**
Does the project involve animals in any way?: **No**
Could the research outputs potentially be harmful to third parties?: **No**
Could your research/artefact be adapted and be misused?: **No**
Does your project or project deliverable have any security implications?: **No**
Please read and confirm that you agree with the following statements: **Confirmed**
Please read and confirm that you agree with the following statements: **Confirmed**
Please read and confirm that you agree with the following statements: **Confirmed**

Supervisor Review

As supervisor, I will ensure that this work will be conducted in an ethical manner in line with the University Ethics Policy.

Supervisor signature: Confirmed by e-mail by Matt Dennis **Date:** 14/09/2020

Appendix C: CITI certificate

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 1 OF 2
COURSEWORK REQUIREMENTS*

* NOTE: Scores on this Requirements Report reflect quiz completions at the time all requirements for the course were met. See list below for details. See separate Transcript Report for more recent quiz scores, including those on optional (supplemental) course elements.

- **Name:** Davide Garofalo (ID: 9024825)
- **Institution Affiliation:** Massachusetts Institute of Technology Affiliates (ID: 1912)
- **Institution Email:** up963994@myport.ac.uk
- **Institution Unit:** School of Computing
- **Phone:** 00393932516344

- **Curriculum Group:** Human Research
- **Course Learner Group:** Data or Specimens Only Research
- **Stage:** Stage 1 - Basic Course

- **Record ID:** 36043815
- **Completion Date:** 29-Mar-2020
- **Expiration Date:** 29-Mar-2023
- **Minimum Passing:** 90
- **Reported Score*:** 97

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
Belmont Report and Its Principles (ID: 1127)	24-Mar-2020	3/3 (100%)
History and Ethics of Human Subjects Research (ID: 498)	29-Mar-2020	5/5 (100%)
Basic Institutional Review Board (IRB) Regulations and Review Process (ID: 2)	29-Mar-2020	5/5 (100%)
Records-Based Research (ID: 5)	29-Mar-2020	3/3 (100%)
Genetic Research in Human Populations (ID: 6)	29-Mar-2020	5/5 (100%)
Populations in Research Requiring Additional Considerations and/or Protections (ID: 16680)	29-Mar-2020	5/5 (100%)
Research and HIPAA Privacy Protections (ID: 14)	29-Mar-2020	4/5 (80%)
Conflicts of Interest in Human Subjects Research (ID: 17464)	29-Mar-2020	5/5 (100%)
Massachusetts Institute of Technology (ID: 1290)	29-Mar-2020	No Quiz

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?k59ac8e21-757f-4c30-bfdc-423c681f8364-36043815

Collaborative Institutional Training Initiative (CITI Program)
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 2 OF 2
COURSEWORK TRANSCRIPT**

** NOTE: Scores on this Transcript Report reflect the most current quiz completions, including quizzes on optional (supplemental) elements of the course. See list below for details. See separate Requirements Report for the reported scores at the time all requirements for the course were met.

- **Name:** Davide Garofalo (ID: 9024825)
- **Institution Affiliation:** Massachusetts Institute of Technology Affiliates (ID: 1912)
- **Institution Email:** up963994@myport.ac.uk
- **Institution Unit:** School of Computing
- **Phone:** 00393932516344

- **Curriculum Group:** Human Research
- **Course Learner Group:** Data or Specimens Only Research
- **Stage:** Stage 1 - Basic Course

- **Record ID:** 36043815
- **Report Date:** 30-Mar-2020
- **Current Score**:** 100

REQUIRED, ELECTIVE, AND SUPPLEMENTAL MODULES	MOST RECENT	SCORE
Basic Institutional Review Board (IRB) Regulations and Review Process (ID: 2)	29-Mar-2020	5/5 (100%)
Belmont Report and Its Principles (ID: 1127)	24-Mar-2020	3/3 (100%)
Records-Based Research (ID: 5)	29-Mar-2020	3/3 (100%)
Genetic Research in Human Populations (ID: 6)	29-Mar-2020	5/5 (100%)
Research and HIPAA Privacy Protections (ID: 14)	30-Mar-2020	5/5 (100%)
History and Ethics of Human Subjects Research (ID: 498)	29-Mar-2020	5/5 (100%)
Populations in Research Requiring Additional Considerations and/or Protections (ID: 16680)	29-Mar-2020	5/5 (100%)
Conflicts of Interest in Human Subjects Research (ID: 17464)	29-Mar-2020	5/5 (100%)
Massachusetts Institute of Technology (ID: 1290)	29-Mar-2020	No Quiz

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?k59ac8e21-757f-4c30-bfdc-423c681f8364-36043815

Collaborative Institutional Training Initiative (CITI Program)
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org