**UNIVERSITY OF CATANIA**
Department of Mathematics and Computer Science
XXXIV Ph.D. in Computer Science

*Paolo Marco Riela*

# Data Augmentation and Machine Learning for Risk Assessment in Healthcare Associated Infections

Doctoral Thesis

Supervisor: Prof. Giovanni Gallo

Academic Period 2018 - 2021

"*Data is a precious thing and will last longer than the systems themselves.*"

Tim Berners-Lee

# *Abstract*

Acquisition and analysis of extensive datasets is, today, a central tool in most research fields. Machine learning provides powerful methods to obtain descriptive and predictive models for the data in many applications. The acquisition of quality information is fundamental for the reliability and accuracy of predictive and classification models increasingly used in various applications. A correct and adequate use of A.I. models integrated with modern visual analytics techniques allows to extend and overcome the classical statistical methods, thus helping experts and professionals of different fields in decisions and policy-making. A key element for the success of machine learning models, beyond the continuous comparison with the experts in the field of application, is represented by the quality and the completeness of the data included in the analysis. The research reported in this Thesis focuses on the analysis, visualization and balancing of data collected in medical studies in the area of healthcare-associated infections (HAIs) to obtain useful classifier models. The results of this interdisciplinary work improve patient risk stratification and lead to targeted infection prevention and control interventions.

This Thesis addresses these two issues with two main contributions: analytics technique designed to display pathways and common patterns in a sequence of events connected to associated outcome and a data augmentation method based on data imputation and oversampling of the minority classes to generate new records for training machine learning models and improve the visual analytics tools. The effectiveness of these methods is proved in selected real-world case studies, allowing to meet the performance requirements of Public Health, in particular with applications of visual analytics methods and machine learning models on medical datasets.

# *Acknowledgements*

I would like to thank my PhD supervisor, prof. Giovanni Gallo for his expert scientific advice and personal guidance in the course of this research.

A personal thank you to my family and friends who supported me during the years I've spent working on this project.

I would also like to thank Tahnee for her loving presence and Ronnie and Susie for their constant support.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

In recent years, large amounts of routinely or automatically collected datasets - which are electronically captured and stored - represent a key component and core element for public health policy-making. In this context, the fusion and connection of existing databases might help to monitor and to evaluate the impact of risk and/or policy changes at population level. Particularly, analysis of big data may contribute to widen information for the prevention of diseases by the identification of risk factors, to improve the effectiveness of interventions, and to predict outcomes [1].

Among the threats to global health, Healthcare-Associated Infections (HAIs) affect millions of patients worldwide, representing one of the main adverse events, especially in Intensive Care Units (ICUs). Since HAIs are an increasing public health problem, surveillance is considered a strategy to ensure health quality [2].

Traditional public health surveillance relies mainly on statistical techniques and it can be improved today with modern visual analytics and A.I. techniques. This poses unique technical challenges such as data sparsity and lack of positive training samples.

This research focuses primarily on the analysis of biomedical and epidemiological data. The aim is to apply and to integrate machine learning and data augmentation methods within state-of-the-art analytical and visual analytics techniques in order to discover and prove new relationships in data. The interdisciplinary Data Science and Public Health approach is a promising way to exceed the limits of traditional statistics and to apply new methods for the analysis of structured and unstructured

data. Finally, transferral of data analysis results to physicians and medical personnel is a crucial issue. The communication gap between data analysts and these professionals may benefit by the application of adequate visual analytics techniques.

## 1.1 Motivation

In biology and medicine, data analysis is an important tool to acquire a better knowledge of the health status of the population and to monitor the efficiency of health services. It is hence of great relevance to provide physicians with user-friendly IT tools and comply with the efficiency standards required by Public Health. For this reason the developing of visual analytics tools integrated with machine learning models is a key element to automate the entire data analysis workflow for providing deeper and more comprehensive insights. The correct and adequate use of A.I. becomes an extension of the common data analysis that helps medical personnel in decisions and policy-making.

HAIs are the most frequent adverse outcome occurring when patients stay in hospital wards, especially in intensive care units (ICUs) and patients admitted to ICUs generally had a worse clinical prognosis, including prolonged hospital stays, sepsis and mortality [3]. Particularly, mortality in ICUs is two times higher among infected patients than those not infected [4, 5].

Nowadays, several early warning scores (EWS) have been proposed as a helpful way to monitor patients clinical deterioration and disease severity during their stay in ICUs [6–9]. In clinical practice, the Simplified Acute Physiology Score (SAPS) II represents the most commonly used score. Specifically, it is able to predict patients prognosis and to estimate their risk of HAIs, sepsis and dying, according to seventeen physiological variables at ICU admission [2, 10–16]. EWS combined with large amount of patients' history data can be used to train machine learning models to obtain a better outcome predictive ability. Unfortunately, real world data are in most cases incomplete and/or imbalanced and it is one of the main causes for the decrease of generalization in machine learning algorithms. This is typical of medical datasets where data acquisition procedures can cause a lot of missing records and high risk patients tend to be the minority class. Therefore, there is a need for

good data augmentation techniques for medical datasets to improve the efficiency of predictive models and visual analytics tools.

With this in mind, the complexity of HAI burden suggested the need of novel approaches aimed at early identifying patients at higher risk of adverse events in ICU [1]. Indeed, the prediction of patients at higher risk of mortality in ICU play a key role in improving patients survival and in implementing their management [17]. Although several traditional statistical approaches are widely used in clinical practice, modern machine learning models have shown more accurate results in the early identification of patients who are more likely to be infected or die during their stay in ICU, considering different sets of risk factors [17–22].

## 1.2 Research outline

The outline of the project follows a quite straightforward paths for this kind of studies: data acquisition, storage, security, anonymisation, warehousing and cleaning. It is completed with the development of data augmentation and visual tools to ease the analysis and reading of information.

Data comes from the SPIN-UTI Network, as described in section 6.1. The SPIN-UTI Network has surveyed approximately 20,000 patients, more than 4300 infections and 5300 micro-organisms. A preliminary work has been hence to integrate all the data into a single dataset.

Once data are gathered into a single large dataset, it was possible to select the interested features for the case studies. The unified dataset, however, contained missing data and it was strongly imbalanced in terms of outcomes of interest. A preliminary study has been performed including the theory of G-computation and Causal Graphs, but the implementation of G-computation for longitudinal data is more complex than in the point treatment setting [23]. The nature of the data under investigation led to implement well-suited methods for parametric visualization, based on Outflow graphs and Sankey diagrams, and data augmentation for replacing missing data and for oversampling the minority classes to generate new records for training machine learning models.

The principal aim for this research is to achieve innovative results at the interdisciplinary confluence between Scientific and Information Visualization, Advanced Data Analysis and Public Health Science. Indeed, some of the very powerful techniques arising in Data Analysis (Bayesian networks and G-computation) have not yet been exploited by the biomedical research for lack of a common language and know-how. I am confident that new graphical and visual techniques way strongly contribute to fill in the gap and lead to a new set of useful tools.

## 1.3 Thesis organisation

The Thesis is structured as follows:

- Chapter 2 shows an overview on Big Data and the *quality* of information with a focus on Public Health field.
- Chapter 3 introduces the G-computation method and the use of Causal graphs in the epidemiologic research
- Chapter 4 introduces the Outflow graphs and describes a visualization encoding including case studies
- Chapter 5 describes the Imputation and Data Augmentation methods optimized for medical data
- Chapter 6 exposes the case studies with the applications of different models and results

## 1.4 Publications

## List of publications

G. Favara, P. M. Riela, A. Maugeri, M. Barchitta, G. Gallo, and A. Agodi. "Risk of Pneumonia and Associated Outcomes in Intensive Care Unit: An Integrated Approach of Visual and Cluster Analysis". In: *2019 IEEE World Congress on Services (SERVICES)*. Vol. 2642-939X. 2019, pp. 289–294. DOI: 10.1109/SERVICES.2019.00083.

M. Barchitta, A. Maugeri, G. Favara, P. M. Riela, C. La Mastra, M. C. La Rosa, R. M. San Lio, G. Gallo, I. Mura, A. Agodi, and SPIN-UTI Network. "Cluster analysis identifies patients at risk of catheter-associated urinary tract infections in intensive care units: findings from the SPIN-UTI Network". In: *Journal of Hospital Infection* 107 (2021), pp. 57–63. ISSN: 0195-6701. DOI: https://doi.org/10.1016/j.jhin.2020.09.030. URL: https://www.sciencedirect.com/science/article/pii/S0195670120304552.

M. Barchitta, A. Maugeri, G. Favara, P. M. Riela, G. Gallo, I. Mura, and A. Agodi. "A machine learning approach to predict healthcare-associated infections at intensive care unit admission: findings from the SPIN-UTI project". In: *Journal of Hospital Infection* 112 (2021), pp. 77–86. ISSN: 0195-6701. DOI: https://doi.org/10.1016/j.jhin.2021.02.025. URL: https://www.sciencedirect.com/science/article/pii/S0195670121000840.

M. Barchitta, A. Maugeri, G. Favara, P. M. Riela, G. Gallo, I. Mura, and A. Agodi. "Early Prediction of Seven-Day Mortality in Intensive Care Unit Using a Machine Learning Model: Results from the SPIN-UTI Project". In: *Journal of Clinical Medicine* 10.5 (2021). ISSN: 2077-0383. DOI: 10.3390/jcm10050992. URL: https://www.mdpi.com/2077-0383/10/5/992.

G. Gallo, F. Buscemi, M. Ferro, M. Figuera, and P. M. Riela. "Abstracting Stone Walls for Visualization and Analysis". In: *Pattern Recognition. ICPR International Workshops and Challenges*. Ed. by A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, and R. Vezzani. Cham: Springer International Publishing, 2021, pp. 215–222. ISBN: 978-3-030-68787-8.

M. Barchitta, A. Maugeri, G. Favara, R. Magnano San Lio, P. M. Riela, L. Guarnera, S. Battiato, and A. Agodi. "Healthy diet and lifestyles assessment using a mEMA approach: protocol of the HEALTHY-UNICT study". In: *European Journal of Public Health* 31 (2021). ISSN: 1101-1262. DOI: 10.1093/eurpub/ckab165.408. URL: https://doi.org/10.1093/eurpub/ckab165.408.

A. Maugeri, M. Barchitta, R. Magnano San Lio, M. La Rosa, G. Favara, L. Guarnera, P. M. Riela, S. Battiato, and A. Agodi. "Design, protocol, and

perspectives of the MADRE-REA study". In: *European Journal of Public Health* 31 (2021). ISSN: 1101-1262. DOI: 10.1093/eurpub/ckab164.035. URL: https://doi.org/10.1093/eurpub/ckab164.035.

# Chapter 2

# Dealing with data

The total amount of data created, captured, copied, and consumed directly by the users or routinely generated and collected by machines is forecast to increase rapidly in next years. The fast development of digitalization contributes to the ever-growing global data sphere.

The term Big Data was used, since the 90s, to indicate large datasets of unstructured, semi-structured and structured data with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time [24].

To manage big data we need to deal with several challenges related to acquisition, storage, cleaning, analysis and visualization. In the last years, data analysis became the main tool to discover hidden value of data. Concepts like predictive analytics and advanced machine learning methods could help to find new relationship in available data.

## 2.1 The Big Data paradigm

Coined in 1997 by two NASA scientists that found difficult to visualize and memorize a huge dataset [25], the term Big Data in 2001 was defined through the model of 3 Vs. For Big Data we mean those data that have one or more of the following features [26]:

- Variety, in terms of format, source and structure (or lack of a real structure)

- Volume, ie large amounts of data (generated automatically by machines, sensors, DCS, scientific instruments or relating to banking transactions and movements on the financial markets)
- Velocity, ie the speed with which data is produced

Different Vs was added in these years: they represented the qualities of big data in volume, variety, velocity, veracity, and value. Variability is often included as an additional quality of big data:

- Veracity, which refers to the data quality and value [27]. It can affect the accuracy of the analysis
- Value: it can be achieved by the analysis and the processing of large datasets
- Variability: formats, structure or sources of data. Big data can include structured, unstructured or semi-structured data.

In 2018 a new definition asserts "Big data is where parallel computing tools are needed to handle data" [28].

## 2.2 Quality and completeness

There are several definitions of data quality that depend on the contexts which data are used in. Defining it in a sentence is not simple, but we can assert that data quality depends on the state of qualitative or quantitative information. Moreover, the data is considered of high quality if they correctly represent the real world to which they refer.

A common problem encountered by machine learning professionals when analyzing real-world information regards the missing data. As many statistical models and machine learning algorithms rely on complete datasets, it is key to handle the missing data appropriately. Moreover, machine learning algorithms generally require large datasets to be trained [29].

### 2.2.1 Data quality in Public Health

The large amounts of information acquired in recent years, especially in the biomedical field both in terms of research and public health (eg. electronic medical records - EMRs), and the different nature of the collected data, of a quantitative type (as test results of laboratory), qualitative (eg. documents and textual demographic data) and transactional (eg. records relating to drugs), have brought more and more significant implications in the epidemiological and public health fields.

However, much of this rich dataset is currently perceived as a byproduct of healthcare delivery, rather than a central asset to improve its efficiency [30].

In this context, the fusion and connection of existing databases might help to monitor and to evaluate the impact of risk and/or policy changes at population level. Particularly, analysis of big data may contribute to widen information for the prevention of diseases by the identification of risk factors, to improve the effectiveness of interventions, and to predict outcomes [31].

As in other fields, missing data are a pervasive problem in many public health investigations. The standard approach is to restrict the analysis to subjects with complete data on the variables involved in the analysis. Estimates from such analysis can be biased, especially if the subjects who are included in the analysis are systematically different from those who were excluded in terms of one or more key variables [32].

In this research we will face different approaches, from the most theoretical to the most applicative ones, for the reconstruction and generation of data in the medical and epidemiological fields. In the case studies dealt with during the course of PhD, the most relevant problems related to data quality concerned a massive lack of data for the selected features and a strong imbalance of the classes for the outcomes of interest. G-computation was explored to obtain parameters from observational data (section 3.1), but a machine learning approach based on k-NN was implemented for data imputation (subsection 5.1.3) and balancing, excluding undersampling techniques, (section 5.2) to generate new data for training models. The original records with a complete set of features was used for testing the models.

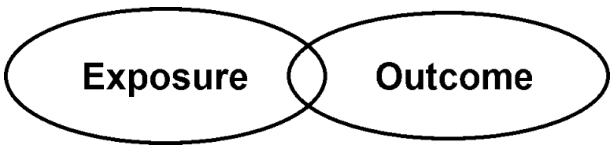
# Chapter 3

# G-computation

Often in causal inference literature different statistical methods are used to estimate causal effect from epidemiological observational data.

Causal inference is the process of determining the independent, actual effect of a particular phenomenon that is a component of a larger system. The main difference between causal inference and inference of association is that causal inference analyzes the response of an effect variable when a cause of the effect variable is changed [33, 34], while inference of association is a statistical relationship between two variables, like a specified health outcome more likely in people with a particular exposure (figure 3.1). In this case two variables may be associated without a causal relationship (the classic example of correlation not equaling causation can be found with ice cream and murder: the rates of violent crime and murder have been known to jump when ice cream sales do).

Figure 3.1: Example of association




## 3.1 Overview

G-computations (G stays for general) was introduced by James Robins in 1986. It is one instance of the so called Marginal Structural Models (MSM). Several works

focused on G-computation are present in the epidemiology literature including theoretical explanation of the method. This paragraph describes an overview of this approach.

The G-computation is a method that allows investigators to use observational data to estimate parameters that in ideal situation should be obtained in a perfectly randomized controlled trial. Under certain assumptions, these estimates can be interpreted causally. G-computation, a maximum likelihood substitution estimator of the G-formula, is one such approach to causal-effect estimation [35].

Let $Y(1)$ and $Y(0)$ be the two potential outcomes under the exposure and the non-exposure, respectively [36]. Let $(Z, X)$ denote the random variables related to the exposure statuses of individuals ($Z = 1$ for exposed individuals and 0 otherwise) and the $k$ covariates ($X = X_1, ..., X_k$) measured before exposure, respectively. The average causal effect is $ACE = E[Y(1) - Y(0)]$. It represents the mean difference between the outcomes of individuals if they had been exposed or unexposed [37].

G-computation is a natural extension of the traditional regression techniques; in fact, the first step of G-computation is a traditional regression. As with other causal inference techniques, the G-computation approach decouples the estimation of effects of interest from the estimation of parameters that are not directly related to the research question (e.g., effects of confounders). Additionally, when the effect of exposure on the outcome varies by strata of a third covariate -in other words, interaction exists for the treatment variable- G-computation permits the estimation of a single, marginal effect estimate averaged across the observed distribution of that third variable. The estimation of a single effect may simplify interpretation of exposure effects as compared with multiple effect estimates, depending on the research question [38].

Unfortunately G-computation is not well-suited to every data structure and the implementation of the method for longitudinal studies is more complex than in the point treatment setting. For this reason in the next chapters will be described machine learning approaches useful to obtain missing or additional information from longitudinal datasets.

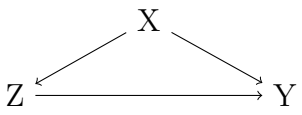
## 3.2 Causal graphs

Causal graphs in the form of path diagrams are an integral component of path analysis [39] and structural equations modeling [40]. The theory of causal graphs is equivalent to the G-computation theory of Robins [41, 42]. It has a benefit, however, of providing a compact graphical as well as algebraic formulation of assumptions and results, which may be easier for the general reader and for the specialized field practitioner to comprehend. In addition, it provides a novel perspective on traditional epidemiologic criteria for confounder identification [43].

A causal graph is a direct acyclic graph (DAG) and it consists of vertices and edges, with each edge directed from one vertex to another, such that following those directions will never form a closed loop.

Fig. 3.2 describes a simple scenario where $Y$ represents the outcome, $Z$ the exposure and $X$ a covariate and potential confounder that distorts the perceived causal relationship between $Z$ and $Y$ if unaccounted for. Confounder is defined as a situation in which the study exposure groups differ in their probability distribution for the outcome for reasons other than effects of exposure [43]. So the marginal association between $Z$ and $Y$ may be partly causal and partly confounded.

Figure 3.2: A simple causal graph




Causal graphs have a long history of formal and informal use and the theory is widely described in the literature. They can provide a starting point for identifying variables that must be measured and controlled to obtain unconfounded effect estimates. They also provide a method for critical evaluation of traditional epidemiologic criteria for confounding [43].

The graphical assumptions of causal graphs are qualitative and nonparametric, in that they imply nothing about the specific functional form of the relations or distributions among the variables. Like in G-computation, this method it is not readily

suitable and easy to implement in studies with a continuous treatment and/or in longitudinal studies with long follow-up with or without time-dependent outcomes.

The data analyzed during the doctoral research concerned hospitalized patients and patients staying in ICU for more than two days (section 6.1). Hospitalization and ICU stay can be described as a series of temporal events. Some events contain significant information that, if put together into event sequences, can reveal insightful facts in patient's history or lead to new relationships in data.

Collections of event sequences are growing rapidly throughout many areas such as electronic medical records (EMRs), sports events, call centers, transportation incident logs, and student progress reports. In addition, many event sequences have associated *outcomes*. [44]. For example, outcomes for ICU stay data could be measured by mortality or discharge rates.

Chapter 4 describes a well-suited DAGs method for time events series and the original contribution for its visualization development.

# Chapter 4

# Quantitative visualization method for Outflow graphs

A first part of the research work has focused on the deepening and use of state-of-the-art scientific visualization methods. In particular, we focused on visualization methods that have been classified under the name of "Visual analytics". Among the many methods available, particular attention was given to the Outflow graphs method and its quantitative graphic coding discussed below.

## 4.1 Outflow graphs

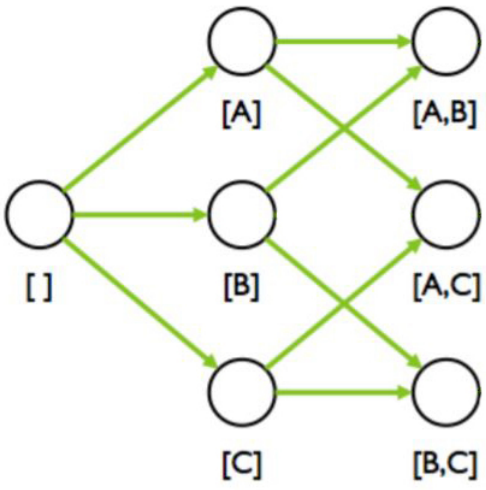



Figure 4.1: Outflow graph

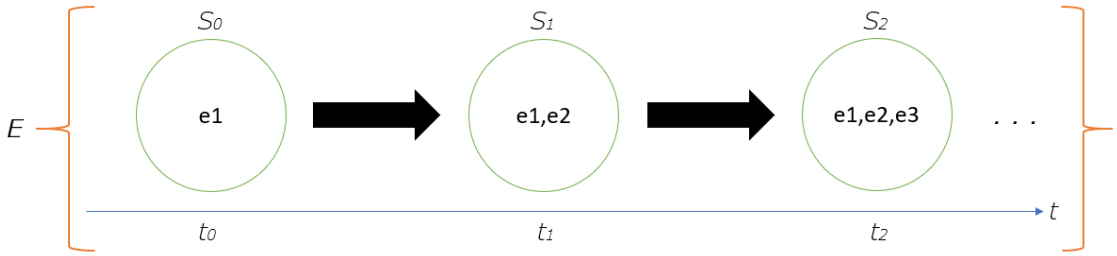
An Outflow graph is a DAG used to visualize event sequences. Collections of discrete events can be aggregated together to form pathways for observing progressive sequences over time. Thus, alternative chains of events may lead to different outcomes.

Fig. 4.1 describes a portion of a simple Outflow graph. The first node (starting point) represents a state without events. It points to nodes with a single event (A, B or C) and after this the edges lead to nodes with a set of aggregated events.

The aggregation of data is a fundamental factor in this type of approach and can be formally defined through the following elements: ***Entities***, ***States***, ***Events*** and ***Outcomes***.

The entity $E$ can be defined as a path that passes through different states $S_i$ ranging from $S_m$ to $S_n$ in the time $T_{m \to n}$. Each state is defined as a set of zero or more events that an entity has encountered before or during time $t_i$ (Fig. 4.2). Each entity ends its path with an Outcome. Different entities can have equal or distinct Outcomes [44].

Figure 4.2: Example of a single Entity




Aggregation of multiple entities occurs under the following assumptions:

1. events are persistent
2. the order of events does not matter

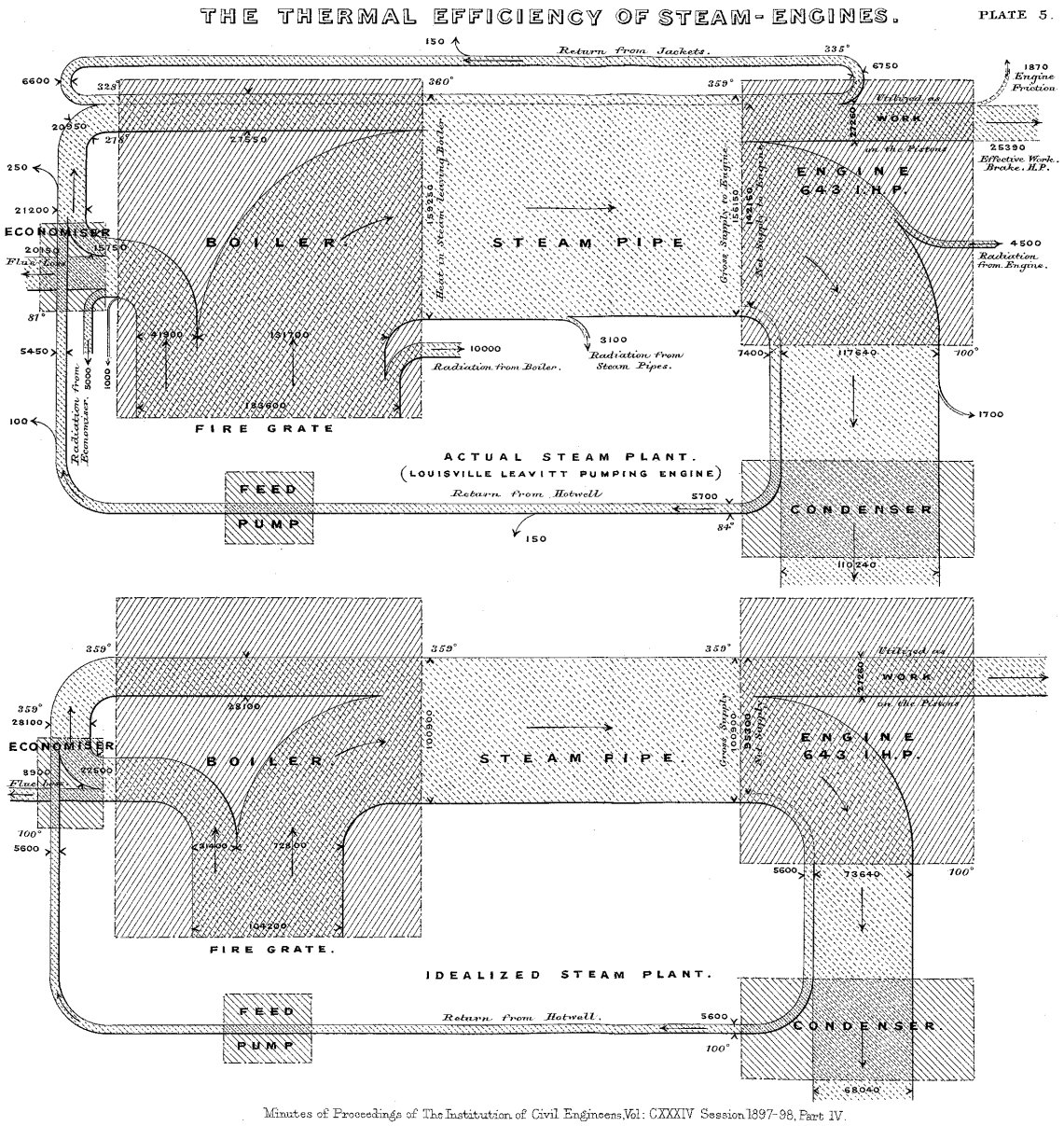
One of the strengths of this method is the independence of the graph dimension from the number of records of a dataset. The only element that weighs in this sense is the number of considered states.

## 4.2 Visual encoding

A more comprehensive visual encoding has been implemented to include quantitative information to the outflow graphs model. The starting idea was inspired by the Outflow Visualization [44] and adapted using the more well-known Sankey diagrams [45].

Sankey diagrams are a type of flow diagram in which the width of the arrows is proportional to the flow rate. They are named after Irish Captain Matthew Henry Phineas Riall Sankey, who used this type of diagram in 1898 showing the energy efficiency of a steam engine [45] (Fig. 4.3).

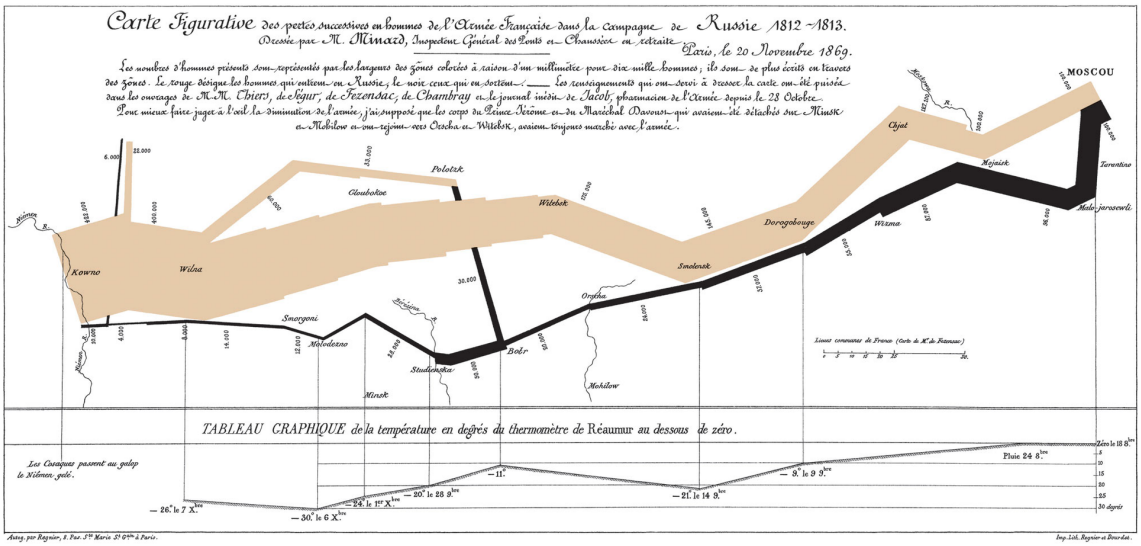
Figure 4.3: Sankey Diagram drawn by M. H. Sankey, extracted from "The Thermal Efficiency Of Steam Engines", showing energy efficiency of a steam engine [46]



One of the most famous Sankey diagrams is Charles Minard's Map of Napoleon's Russian Campaign of 1812. It is a flow map, overlaying a Sankey diagram onto a geographical map. It was created in 1869, predating first Sankey diagram of 1898 (Fig. 4.4).

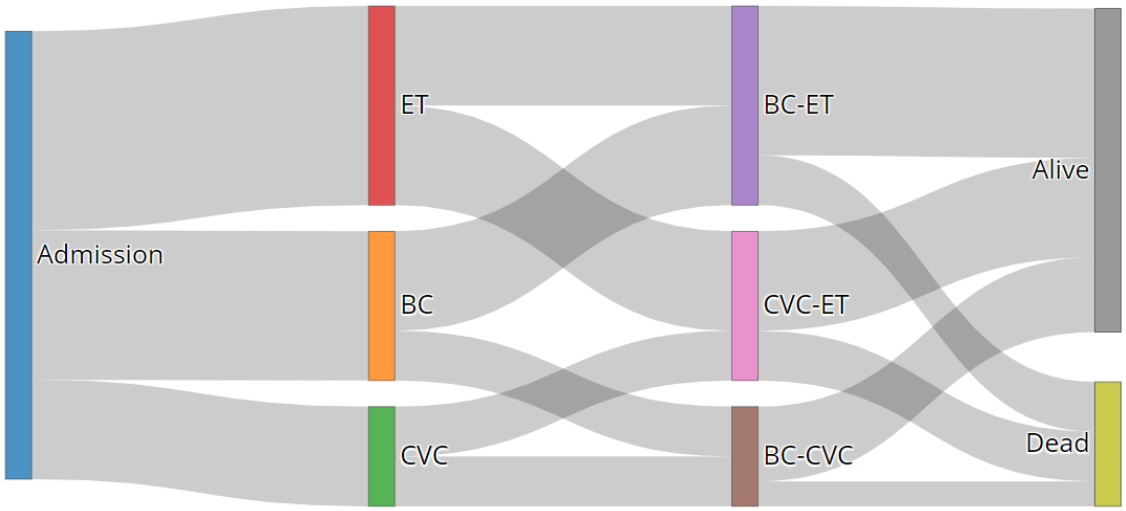
Over time, Sankey diagrams have become a standard model used in science and engineering for representing heat balance, energy flows, material flows, and since the 1990s this visual model has been used in product life cycle assessment [47].

Figure 4.4: Minard's classic diagram of Napoleon's invasion of Russia




A modern version of them is represented by rectangular blocks (nodes) and flows (edges). Each block represents a set, with dimensions proportional to the cardinality of the set, and the flows show the portion of elements that pass from a block to another. The example in figure 4.5 shows a diagram that represents data collected into ICU by the SPIN-UTI network related to a cohort of patients from the moment of admission to the two possible outcomes, discharge from ICU (alive) or death, considering, in order, the application of two different invasive devices to the patients during the stay.

Figure 4.5: Sankey diagram of a cohort of patients admitted in ICU. The blocks represent the status of the patients, from Admission (device free), passing through the application of one and two different invasive devices (ET: endotracheal tube - BD: bladder catheter - CVC: central venous catheter), to reach one of two possible outcomes (Dead or Alive)

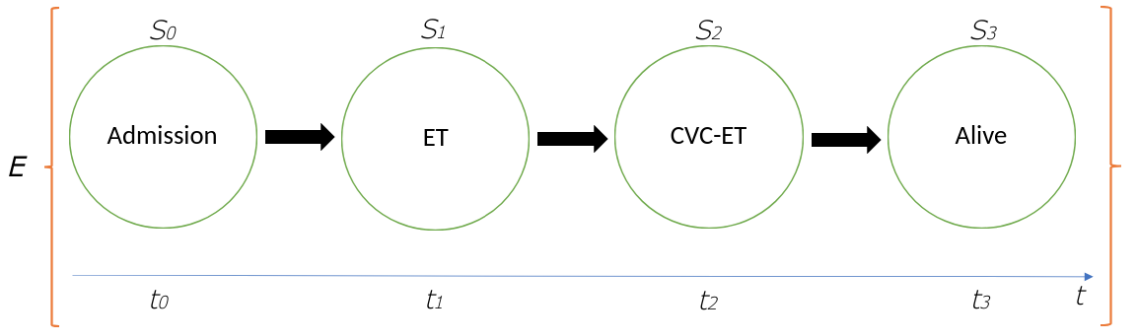



The diagram starts with the admission (device free) state and ends with one of the two possible outcomes (dead/alive). The several subsets of invasive device used on a patient are intermediate nodes of the diagram (ET: endotracheal tube - BD: bladder catheter - CVC: central venous catheter). The width of each edge represents the average number of patients passing from a state to another (eg. from one to two devices) or from a state to the outcome.

These kinds of diagram are interactive and provide information, like the number of elements, about nodes and edges. The application was written in Python + SciPy stack and Plotly library.

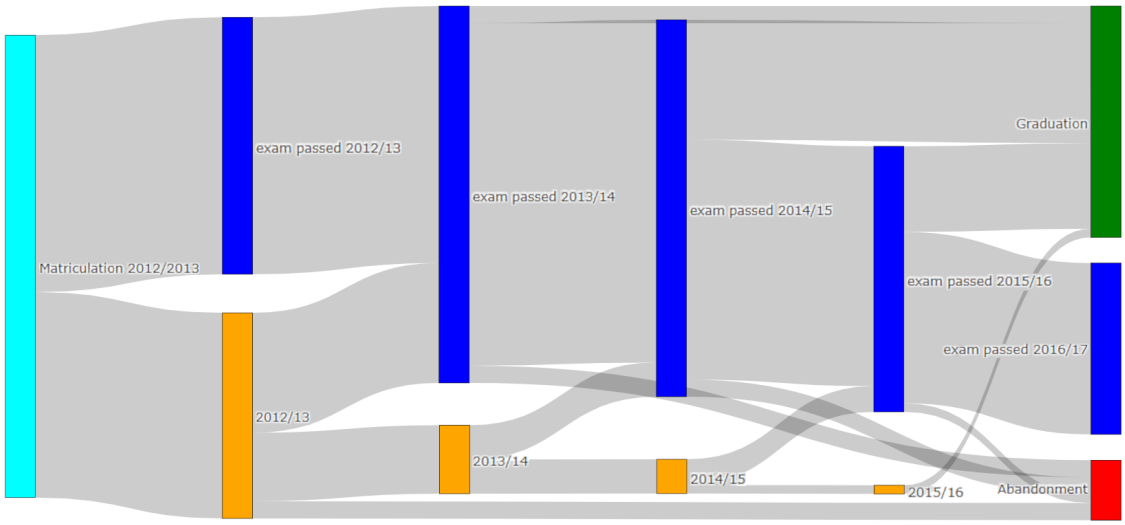
The Sankey diagram in Figure 4.5 is the result of an Outflow graph where the application on a patient of an invasive device is considered as discrete and persistent event. The states are hence defined as subsets of the invasive device used to several stage of ICU stay (Fig. 4.6).

Figure 4.6: A pathway (entity $E$) of patients' ICU stay from state $S_0$ to outcome $S_3$




A different, non-medical, application of this method of visual analytics is shown in the diagram in figure 4.7 which describes a report on the progress of a cohort of students enrolled in a specified academic year (University of Catania - Computer Science, a.y. 2012-2013) in relation to an university exam (Discrete Mathematics).

Figure 4.7: Diagram of student career on single university exam




The diagram analyzes the progression of academic studies for a three-year degree within a time range of 5 years. From the Matriculation state the diagram flows through the different years and show how many students pass a selected exam or not, to reach the three possible outcomes. The outcomes are represented by the Graduation (success), the Abandonment of the degree course (fail) or the continuation outside prescribed time.

From this point of view it is possible to observe which exams are more difficult for students. In this example, about half of the enrolled students pass the specific exam in the first year, while a small part of those who do not pass it, probably discouraged, abandons the studies (first orange block, edge directed to abandonment). Instead, it can be easily seen how a small part of students pass the exam before graduation outside prescribed time: this means that a small percentage of students "drag" the exam to the end of degree course taking longer than expected to graduate.

In conclusion, patient history or student career, and in general "life", can often be described as a series of temporal events. This visual analytics method explores pathways and common patterns in a sequence of aggregated events. Connecting the pathways to the associated outcomes can help data analysts and professionals to better understand the meaning or behavior in the visualized information and to discover how certain paths may lead to different results.

### 4.2.1 Applications

Among the threats to global health, Healthcare-Associated Infections (HAIs) affect millions of patients worldwide, representing one of the main adverse events, especially in Intensive Care Units (ICUs). In Europe, approximately 90.000 hospitalized patients have at least one HAI on a given day [48], which determines a significant increase in mortality and morbidity rates, contributing to the raise in hospital assistance costs. Since HAIs are an increasing public health problem, surveillance is considered a strategy to ensure health quality [49].

The data used in the following case studies comes from the SPIN-UTI network, described in section 6.1.

#### Risk of Pneumonia in ICU

Pneumonia is the most frequent HAI, especially in ICU, where a significant percentage of patients is exposed to mechanical ventilation [50]. In Europe, pneumonia occurs in 7% of patients hospitalized for at least two days in the ICU, among these 91% are associated with mechanical ventilation. Except for the intrinsic patient characteristics, several factors - i.e. the management of intubation procedures and

intubated patients- have been proposed as a potential target to control IAP incidence and associated outcomes [51].

The use of Outflow-like approach to these studies is naturally suggested by the nature of the phenomena under investigation: a patient is considered in an initial state that evolves through the hospitalisation into other states because new events are occurred.

Unfortunately the direct usage of Sankey diagram does not give information about the average time to go from a state to the next in addiction to the non-persistence of some events. For example, the endotracheal tube can be applied several times during a patients stay. This means that the patient passes from the intubated to the non-intubated state and then goes back to the intubated state.

To bypass these issues we have considered the total time of an invasive device applied to 9656 patients enrolled in 92 ICUs of 62 hospitals.

In this study we distinguish three clusters of patients with similar characteristics (Table 4.1). In order to identify them, the following variables were standardized and imputed in the cluster analysis: age, SAPS II score at admission, patient origin and administration of antibiotics within 48 hours of admission [1]. The cluster analysis was conducted in SPSS, using the TwoStep Clustering method (optimal number of clusters selected automatically by the clustering algorithm based on Schwarz's Bayesian Information Criterion), an exploratory tool designed to reveal natural clusters within a dataset that would otherwise not be apparent [52]. The software of choice was useful for medical personnel to visualize and easily handle large datasets thanks to its GUI.

The current analysis was performed on patients staying in ICU for more than 2 days, without missing values in information related to patient characteristics (e.g., age, sex, Simplified Acute Physiology Score II - SAPS II - at admission, origin of patient, admission type), dates of insertion and removal of intubation, and outcomes (e.g., pneumonia, sepsis, death).

Table 4.1: Comparison of population characteristics by clusters

| **Characteristics** | **Cluster 1 (N=2143)** | **Cluster 2 (N=5854)** | **Cluster 3 (N=1659)** | ***p-value*** |
|---|---|---|---|---|
| Age, years | 69.0 (24.0) | 70.0 (20.0) | 70.0 (20.0) | 0.028 |
| Sex (% men) | 62.8% | 61.0% | 60.5% | 0.263 |
| **Origin** | | | | |
| Other ward of this/other hospital | 39.6% | 82.0% | 82.2% | <0.001 |
| Other ICU | 1.1% | 3.7% | 2.7% | |
| Community (home) | 58.5% | 12.7% | 13.2% | |
| Long-term care facility | 0.7% | 1.5% | 1.9% | |
| SAPS II score at admission | 40.0(27.0) | 38.0 (26.0) | 37.0 (23.0) | <0.001 |
| **Type of ICU admission** | | | | |
| Medical | 63.2% | 47.8% | 52.8% | <0.001 |
| Scheduled surgery | 18.8% | 35.6% | 36.6% | |
| Unscheduled surgery | 18.0% | 16.7% | 10.6% | |
| Trauma | 5.7% | 4.4% | 4.4% | 0.043 |
| Impaired immunity | 5.8% | 7.4% | 3.6% | <0.001 |
| Administration of antibiotics within 48 hours of admission | 67.9% | 87.0% | 32.9% | <0.001 |
| Length of stay in ICU, days | 5.0 (10.0) | 5.0 (9.0) | 4.0 (8.0) | 0.134 |

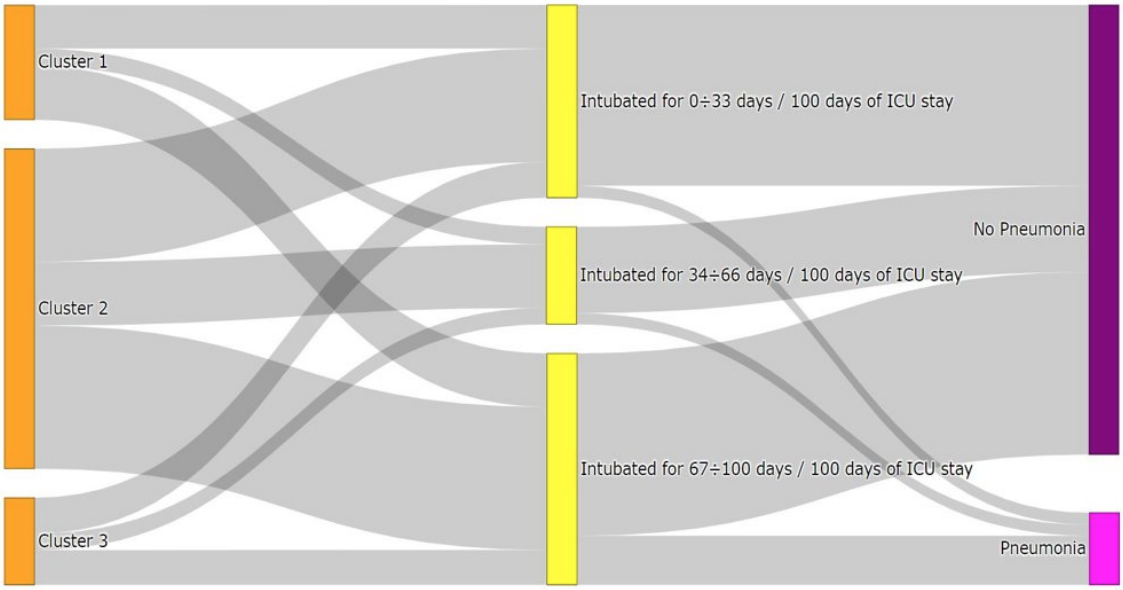
Results are reported as median (interquartile range) for continuous variables, or percentage for bivariate or categorical variables. Statistical analyses were performed using the Kruskal-Wallis or the Chi-squared test.

Let us consider as initial states, in a case, the three patients' clusters (Fig. 4.8) and, in the other case, the three most observed micro-organisms in the patients with pneumonia (Fig. 4.9). Eventually an outcome is reached.

Sankey diagram reported in Figure 4.8 provides the flow of patients from their admission to ICU and visualizes how each cluster and duration of intubation contribute to the diagnosis of pneumonia. Using statistical analysis, we first observed that patients belonging to cluster 1 and 2 had higher duration of intubation in days ($median = 3$, $IQR = 9$ and $median = 3$, $IQR = 8$, respectively) and in days/100 days of ICU stay ($median = 66.7$, $IQR = 102.7$ and $median = 70.0$, $IQR = 75.0$, respectively) than those in cluster 3 ($median = 2$, $IQR = 7$, $p\text{-}trend < 0.001$ and

*median* = 60.1 on 100 days of ICU stay, *IQR* = 100.0 on 100 days of ICU stay, *p-trend* = 0.001, respectively). This was in line with information displayed in Figure 4.8, which showed that a higher percentage of patients belonging to cluster 1 (49.0%) or cluster 2 (50.4%) were intubated for more than 66 days/100 days of ICU stay, compared to cluster 3 (45.0%; *p-trend* < 0.001). Figure 4.8 also help us to graphically report that patients who were intubated for more than 66 days/100 days of ICU stay had higher incidence of pneumonia (11.4%) than those who were intubated for 34-66 or 0-33 days /100 days of ICU stay (5.6% and 2.9%, respectively; *p-trend* < 0.001).

Figure 4.8: Outflow of three clusters of patients from ICU admission to diagnosis of pneumonia. Clusters defined in table 4.1 (TwoStep Clustering method, optimal number of clusters by Schwarzs Bayesian Information Criterion)

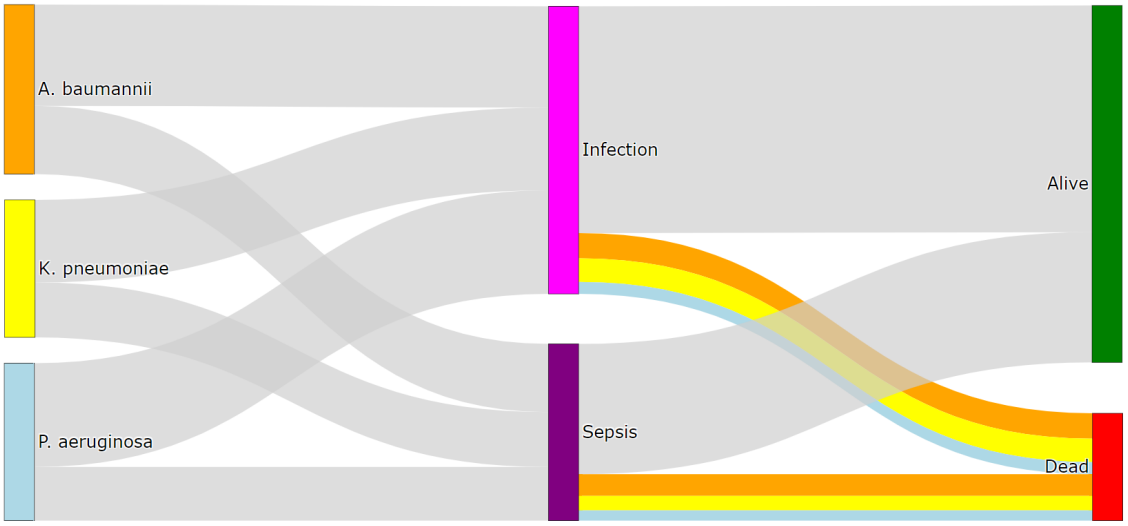



Sankey diagrams are useful for representing this type of data because they allow a visual approach to examine the evolution of patients' sub-sets maintaining the total anonymity of the information. The focus on a specific infection is useful for selecting the fundamental factors and reducing the variables to obtain as result easy-to-read diagrams.

An additional aim of our study was to evaluate risk of sepsis and death associated with pneumonia caused by *Acinetobacter baumannii*, *Klebsiella pneumoniae*

and *Pseudomonas aeruginosa*, the three major causes of pneumonia in our study. Thus, we report a Sankey diagram (Fig. 4.9) illustrating how *Acinetobacter baumannii*, *Klebsiella pneumoniae* and *Pseudomonas aeruginosa*-associated pneumonia contribute to sepsis and mortality. This graphical representation highlighted that patients with *Acinetobacter baumannii* or *Klebsiella pneumoniae*-associated pneumonia were more prone to exhibit sepsis than those infected by Pseudomonas aeruginosa. This finding was further confirmed by statistical analysis, which found a higher incidence of sepsis in patients with *Acinetobacter baumannii* or *Klebsiella pneumoniae*-associated pneumonia (38.9% and 38.8%, respectively) than in those infected by *Pseudomonas aeruginosa* (29.1%; $p = 0.025$). Figure 4.9 also shows that sepsis, in turn, was associated with higher mortality (45.6%) than infection (32.2%; $p < 0.001$). In line with these findings, mortality was higher in patients with *Acinetobacter baumannii* and *Klebsiella pneumoniae*-associated pneumonia (20.6% and 29.1%, respectively) than in those infected by *Pseudomonas aeruginosa* (13.4%; $p < 0.001$).

Figure 4.9: Association of Acinetobacter baumannii, Klebsiella pneumoniae and Pseudomonas aeruginosa to pneumonia and sepsis, and contribution of individual pathogens to death (colored edges directed to "Dead" outcome)




This kind of visual analytics of data, supported by traditional statistical methods, represents a useful instrument to identify and to describe significant determinants associated with adverse outcomes in healthcare. Particularly, our study indicates

that Sankey diagrams are useful tools to visualise flows of patients from their admission to ICU and how each cluster and duration of intubation contribute to the diagnosis of pneumonia. Moreover, they enabled us to graphically represent the contributions of *Acinetobacter baumannii*, *Klebsiella pneumoniae* and *Pseudomonas aeruginosa*-associated pneumonia to the risk of sepsis and death.

**Risk of catheter-associated urinary tract infections in ICU**

Urinary tract infections (UTIs) are among the most common HAIs, representing up to 40% of all HAIs [53]. The presence of a urinary catheter and the duration of exposure allowing continuous access of organisms into the urinary bladder are the main risk factors for development of a catheter-associated UTI (CAUTI) [54]. Indeed, as reported by ECDC, the urinary catheter utilization rate was 78 per 100 patient-days in ICUs, and nearly 98% of UTIs were associated with the presence of a urinary catheter [53]. However, other host factors (i.e. anatomical or functional abnormalities, female sex, older age, diabetes mellitus, genetic predisposition), and bacterial (i.e. pathogen virulence characteristics) and healthcare (i.e. poor quality of catheter care, lack of antimicrobial therapy) characteristics may affect the risk of CAUTIs [55]. The burden of CAUTIs is associated with increased morbidity and mortality, longer length of stay and higher healthcare costs [56]. For instance, in the USA, it has been estimated that CAUTIs cause approximately US$131 million in annual excess medical costs [57]. In addition, urinary catheters are often reservoirs for multi-drug-resistant bacteria and a source of transmission to other patients [58]. CAUTIs are also associated with severe health outcomes including sepsis, a systemic inflammatory condition that occurs when bacteria infecting the urinary tract infect the bloodstream [56]. Surveillance data indicated that sepsis was associated with increased mortality and morbidity in patients of all ages [13, 59]. Although preventive strategies, such as educational initiatives, catheter avoidance and limiting catheter days, have been proposed [60], more efforts are needed to control the incidence of CAUTIs and to improve patient outcomes. In fact, it has been estimated that up to 70% of CAUTIs may be preventable with recommended infection control measures [54, 61–66].

Also in this study we used cluster analysis to distinguish patients according to their characteristics at ICU admission, and to identify clusters of patients at higher

risk for CAUTIs and associated sepsis. Accordingly, variability across clusters in terms of duration of urinary catheterization, and incidence of CAUTIs and associated sepsis was explored [31]. In particular, the two-step clustering method was performed to identify different clusters of patients based on age, sex, SAPS II score at admission, patient origin, type of admission, trauma, and administration of antibiotics in 48 h before or after ICU admission [1]. The clustering algorithm, based on Schwarz's Bayesian Information Criterion (SBIC), allowed sets of clustered variables to be categorized.

The cluster solution obtained was tested by excluding variables with predictive importance $< 0.2$.

The original dataset was built by recording data related to ICU characteristics (type, percentage of mortality, proportion of intubated patients, proportion of patients with a urinary catheter), patient characteristics at admission (e.g. age, sex, SAPS II score, patient origin, admission type), dates of insertion and removal of invasive devices (e.g. urinary catheter), infection status (i.e. infection date, infection site, associated micro-organisms) and micro-organisms (i.e. antimicrobial resistance data).

Table 4.2 shows the characteristics of participants with relative within-cluster homogeneity and between-cluster variability in terms of age,sex, SAPS II score at admission, patient origin, type of admission, trauma, and administration of antibiotics in 48 h before or after ICU admission. In particular, Cluster 1 ($N = 2143$) comprised more patients with a medical type of ICU admission who came from the community. This cluster was also characterized by an intermediate percentage of patients who received antibiotics in 48 h before or after ICU admission, higher proportion of trauma patients, lower median age and higher SAPS II score. Cluster 2 ($N = 5854$) consisted of patients who were more likely to come from other wards/hospitals, and to report administration of antibiotics 48 h before or after ICU admission. This cluster included older patients with an intermediate SAPS II score, and approximately half of them reported a surgical type of ICU admission (i.e. 52.2%, the highest percentage across clusters). Patients in Cluster 3 ($N = 1659$) were similar to those in Cluster 2 in terms of patient origin, type of admission and age. However, Cluster 3 was characterized by a lower percentage of patients with

administration of antibiotics 48 h before or after ICU admission, and lower SAPS II score. No difference in terms of sex distribution across clusters was evident [31].

Table 4.2: Characteristics of clusters of patients at intensive care unit (ICU) admission and urinary catheter utilization

| **Characteristics** | **Cluster 1 (N=2143)** | **Cluster 2 (N=5854)** | **Cluster 3 (N=1659)** | ***p-value*** |
|---|---|---|---|---|
| Age, years | 69 (24) | 70 (20) | 70 (20) | 0.028 |
| Sex (% men) | 62.8% | 61.0% | 60.5% | 0.263 |
| **Patient origin** | | | | |
| Other ward/healthcare facility | 41.5% | 87.3% | 86.8% | <0.001 |
| Community | 58.5% | 12.7% | 13.2% | |
| SAPS II score at admission | 40 (27) | 38 (26) | 37 (23) | <0.001 |
| **Type of ICU admission** | | | | |
| Medical | 63.2% | 47.8% | 52.8% | <0.001 |
| Surgical | 36.8% | 52.2% | 47.2% | |
| Trauma | 5.7% | 4.4% | 4.4% | 0.043 |
| Impaired immunity | 5.8% | 7.4% | 3.6% | <0.001 |
| Antibiotic treatment in 48 h before or after ICU admission | 67.9% | 87.0% | 32.9% | <0.001 |
| Length of ICU stay, days | 5 (10) | 5 (9) | 4 (8) | 0.134 |
| Presence of urinary catheter during ICU stay | 82.9% | 84.1% | 85.6% | <0.001 |
| Duration of urinary catheterization, days | 7 (12) | 7 (11) | 6 (8) | <0.001 |

Results are reported as median (interquartile range) for continuous variables, or percentage for categorical variables. Statistical analyses were performed using the KruskalWallis or the Chi-squared test.

Although length of ICU stay was similar across clusters, visual inspection through Outflow-like approach and Sankey diagrams encoding (Fig. 4.10) revealed differences in terms of urinary catherization and its duration. In fact,participants belonging to Clusters 1 or 2 were less likely to be catheterized (82.9% and 84.1%, respectively) than patients in Cluster 3 (85.6%; $p < 0.001$). However, patients in Clusters 1 or 2 had a longer duration of urinary catheterization (median 7 days, IQR 12 days for

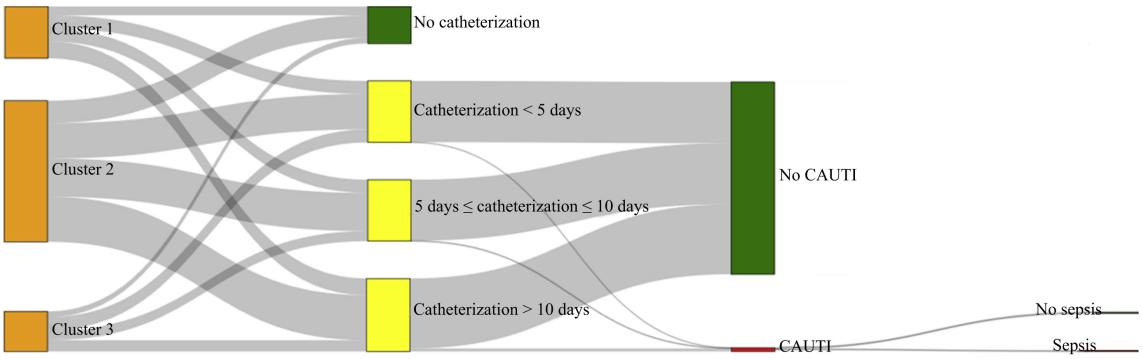
Cluster 1; median 7 days, IQR 11 days for Cluster 2) compared with patients in Cluster 3 (median 6 days, IQR 8 days; $p < 0.001$).

This visual encoding results useful again to visualize the the flow of patients during their ICU stay, without detailed temporal information, helping researchers and medical personnel to reveal and to describe significant factors associated to the averse outcomes.

In general, patients with urinary catheterization exhibited a higher incidence of UTIs than patients who were not catheterized (3.0 per 100 patients vs 1.2 per 100 patients; $P = 0.004$). The rate of CAUTIs was 3.2 per 1000 catheter-days, with an incidence that increased with increasing duration of catheterization: 0.4 per 100 patients in those catheterized for $< 5$ days, 0.8 per 100 patients in those catheterized for $\geqslant 5$ days and $\leqslant 10$ days, and 7.2 per 100 patients in those catheterized for $> 10$ days ($P < 0.001$). Interestingly, patients in Cluster 1 showed a higher incidence of CAUTIs (3.5 per 100 patients) than those in Clusters 2 or 3 (2.5 per 100 patients in both clusters; $P = 0.033$).

Finally, this study found that 37.0% of patients with CAUTIs developed sepsis, but no difference was evident in the incidence of sepsis across clusters (P=0.238). However, the percentage of sepsis among patients with CAUTIs increased with increasing duration of catheterization: 30.0% in participants catheterized for $<$ 5 days, 35.0% in those catheterized for $>$ 5 days and $<$ 10 days, and 45.6% in those catheterized for $>$ 10 days ($P = 0.010$).

Figure 4.10: Outflow of three clusters of patients during ICU stay to diagnosis of CAUTI and sepsi. Clusters defined in table 4.2 (TwoStep Clustering method, optimal number of clusters by Schwarzs Bayesian Information Criterion)



# Chapter 5

# Synthetic data generation with Machine Learning

Fragmented datasets can have a major impact on performance and quality of machine learning algorithms. A small number of missing values (NA) within a large dataset can be bypassed by deleting incomplete records: in most cases this does not lead to a serious loss of information. If the missing data represent a large slice of the dataset, dropping incomplete records involves a heavy loss of information, impacting considerably on statistical and machine learning models.

Furthermore, most classification algorithms, in general, require a balanced dataset in terms of the outcome of interest. An unbalanced dataset often provides acceptable levels of *accuracy*, but with poor results in terms of *f1-score*.

## 5.1 Imputation techniques

There are different techniques in literature to approach the missing data issue. A common way is to ignore it, but this is not suitable when the data are too fragmented. A widely used approach is imputation. Imputation simply means replacing the missing values with an estimate, then analyzing the full dataset as if the imputed values were actual observed values.

Common techniques consist in the replace of NA values with 0, mean, median or mode values. These methods are simple and easy to implement but are often a potential cause of data bias.

Another commonly used method is the regression imputation. In this kind of imputation the predicted value is obtained by regressing the missing variable on other

variables. This preserves relationships among variables involved in the imputation model, but not variability around predicted values.

### 5.1.1 Types of missing data

The missing data issue can be very complex since it is necessary to distinguish the univariate from the multivariate missing data and the forms of missingness that take different types, with different impacts on the validity of the results of research: *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR) [67].

To formally define missingness, lets consider, for simplicity, the univariate case. A dataset $X$ can be divided in two parts:

$$X = \{X_o, X_m\}$$

where $X_o$ corresponds to the observed data, and $X_m$ to the missing data in the dataset.

For each observation we define a binary response whether or not that observation is missing:

$$R = \begin{cases} 1 & if \;\; X\, observed \\ 0 & if \;\; X\, missing \end{cases}$$

The missing value mechanism can be understood in terms of the probability that an observation is missing $\Pr(R)$ given the observed and missing observations, in the form:

$$\Pr(R|x_o, x_m) \text{ [68]}$$

The three mechanisms are subject to whether the probability of response $R$ depends or not on the observed and/or missing values:

Values are **MCAR** if the events that lead to any particular data-item being missing are independent both of observable variables and of unobservable parameters of interest, and occur entirely at random [69]. In this case the probability of an

observation being missing depends only on itself, and reduces to $\Pr(R|x_o, x_m) = \Pr(R)$. As an example, imagine that a doctor forgets to record the gender of every six patients that enter the ICU. There is no hidden mechanism related to any variable and it does not depend on any characteristic of the patients [68]. When data are MCAR, the analysis performed on the data is unbiased; however, data are rarely MCAR.

**MAR** occurs when the missingness is not random, but where missingness can be fully accounted for by variables where there is complete information. When data are MAR, the missingness of data is systematically related to the observed but not the unobserved data [70]. This case is not completely random, but it is the most general case where we can ignore the missing mechanism, as we control the information upon which the missingness depends, the observed data. Said otherwise, the probability that some data is missing for a particular variable does not depend on the values of that variable, after adjusting for observed values. Mathematically the probability of missing reduces to $\Pr(R|x_o, x_m) = \Pr(R|x_o)$. Imagine that if elderly people are less likely to inform the doctor that they had a pneumonia before, the response rate of the variable pneumonia will depend on the variable age [68]. Analyses of datasets containing MAR data may or may not result in bias.

**MNAR** is data that is neither MAR nor MCAR (i.e. the value of the variable that's missing is related to the reason it's missing) [69]. When data are MNAR, the missingness of data is systematically related to the unobserved data: it depends on events or factors which are not measured by the researcher. Determining the missing mechanism is usually impossible, as it depends on unseen data. For example, we can imagine that patients with low blood pressure are more likely to have their blood pressure measured less frequently (the missing data for the variable blood pressure partially depends on the values of the blood pressure) [68]. As with MAR data, complete case analysis of a dataset containing MNAR data may or may not result in bias.

Variables with different types of missing data should be treated separately.

### 5.1.2 Choice of imputation method

Different imputation methods are expected to perform differently on various datasets. There are generic methods in literature to evaluate the performance of various imputation methods on incomplete datasets, in order to help selecting the most appropriate method.

In this research, the imputation method is also linked to the responsiveness of the balancing method that will be described below and to the performance of the machine learning model chosen for the prediction of the outcomes.

The approach is described in the case studies in the next chapter in which the performance of a predictive model is tested on the datasets completed by imputation and balancing methods.

### 5.1.3 k-NN imputation

The nature of the information examined during this research, data with different types of missing values, rich in continuous and categorical variables, in particular dichotomous ones, which do not follow known distributions or probabilistic models, led us to the use of a k-NN imputation method to reconstruct part of the missing values, according to Malarvizhi and Thanamani [71], and adapted to our needs. The k-NN is a non-parametric and lazy learning algorithm which means there is no assumption for underlying data distribution.

The k-NN method is useful for matching points with their closest k-neighbors in a multi-dimensional space and it can be applied on continuous, discrete, ordinal and categorical data which makes it particularly useful for dealing with all kind of missing values. The k-NN method is based on the assumption that a point value can be approximated by the values of the points that are closest to it, based on the other variables. The similarity of two points is determined using a distance function which can be Euclidean, Manhattan, Mahalanobis, etc.

In particular, in our data, we found very effective the use of Jaccard distance in reconstructing the missing dichotomous data in a coherent way. The Jaccard

distance $d_j$ is complementary to the Jaccard coefficient $J$, defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$0 \leqslant J(A,B) \leqslant 1$$

$$d_j(A,B) = 1 - J(A,B)$$

The advantages of the k-NN algorithm are that it considers the correlation structure of the data and predicts with a good accuracy the conditional probability distribution around an observation making properly informed estimations.

In the k-NN algorithm, a higher value of $k$ would include attributes which are significantly different from our target observation, while lower value of $k$ can implies overfitting and increase the variance, and an high variance can increase the influence of random noise in the data. But with a lower $k$ the bias is low: fitting a model to the *1-nearest* point means that the model will be close to the data. In the data imputation point of view, the objective is to reconstruct the missing data as much as similar to the real ones. High bias can cause the algorithm to miss the relations between features and target variables.

In the case studies described in the next chapter, to reconstruct part of the incomplete records, we have used a 1-NN imputation to fill the missing values on each variable. We have applied the algorithm to the dataset divided by outcomes of interest for every different target variable, considering Euclidean distance for non-binary variables and Jaccard distance for dichotomic variables.

The algorithm 1 describes the process of data imputation on a dataset $D_s$ with missing values considering the binary outcome $O$. At the end of the procedure it returns the reconstructed data for class 0 and 1. The process can be repeated on the entire dataset with the reconstructed data.

**Algorithm 1** Multiple data imputation

$D_S$: the dataset with missing data
$O$: the outcome variable in $D_S$
$kNN$: the k-NN imputer with $k = 1$ and $metrics = (Euclidean, Jaccard)$
**procedure** DATAIMPUTATION($D_S$, $O$)
    $m_{O_0} \leftarrow$ list of variables for class 0
    $m_{O_1} \leftarrow$ list of variables for class 1
    **for each** variable $v_0$ in $m_{O_0}$ **do**
        $I_0 \leftarrow kNN(D_S[v_0])$
    **end for each**
    **for each** variable $v_1$ in $m_{O_1}$ **do**
        $I_1 \leftarrow kNN(D_S[v_1])$
    **end for each**
    **return** $(I_0, I_1)$
**end procedure**

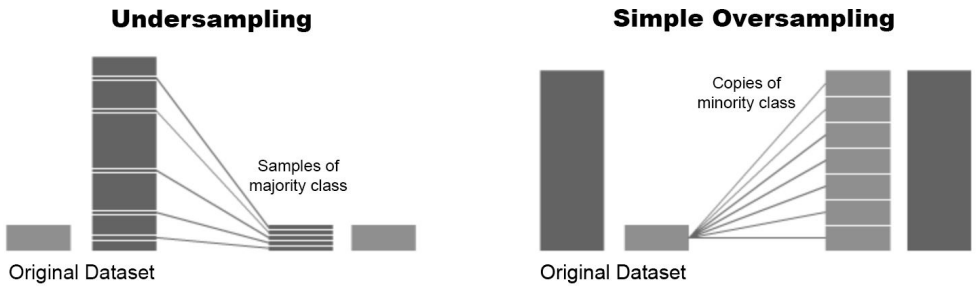
## 5.2 Data balancing methods

In general, classification algorithms require a balanced dataset in terms of the outcome of interest. Several techniques (i.e. undersampling, and oversampling) are currently used to balance the dataset by replicating data or by generating synthetic data.

The **undersampling** method consists to reduce the majority class, keeping only a part of its records. Generally, some observations are randomly eliminated from the majority class in order to match the numbers with the minority class (Fig. 5.1). More advanced undersampling techniques have been used in the past (e.g. undersampling specific samples like the ones further away from the decision boundary [72]), but they did not bring any improvement than simply selecting samples at random. The end result is the same: a smaller number of rows in the dataset. Undersampling is not recommended when the minority class it too small because it heavily reduce the dataset, with a loss of potentially useful information, thus giving the model less data for training so the model is more prone to error.

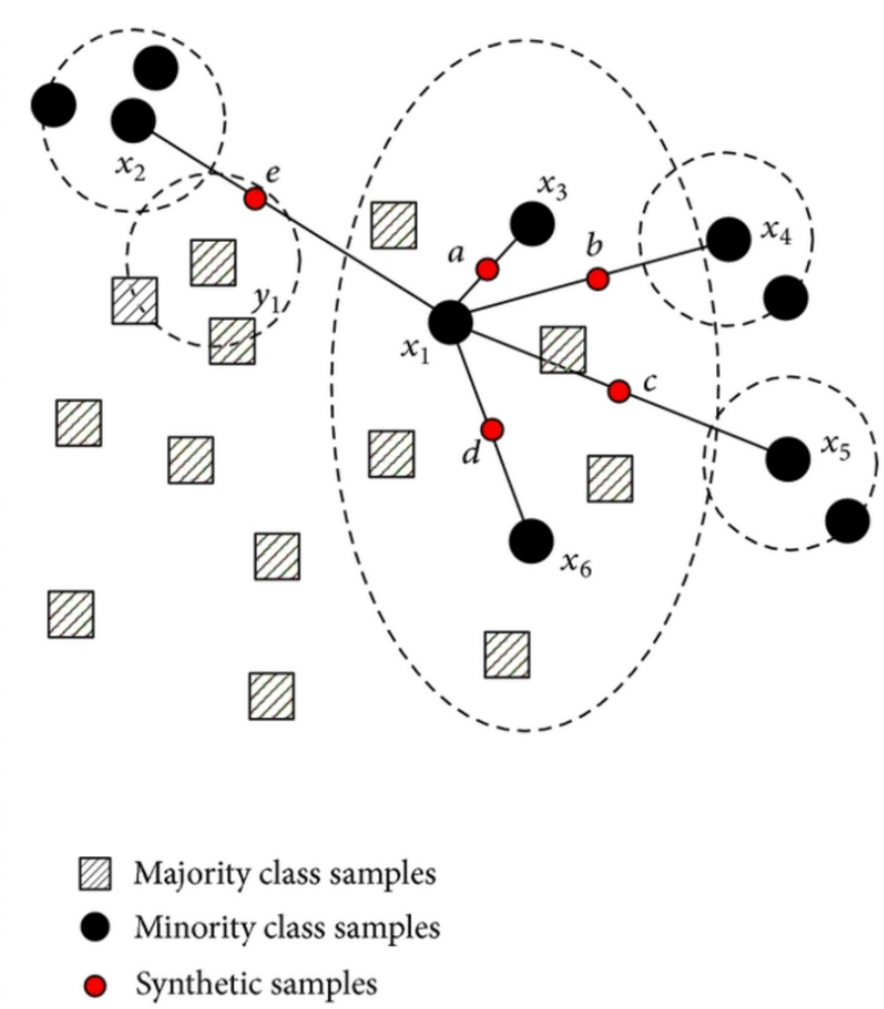
The **oversampling** method, instead, is used to increase the cardinality of the minority class. The easiest way to oversample is to re-sample the minority class duplicating the observations (Fig. 5.1). Thus, the resampled data are exactly the same already present in the dataset. Training a model with too much duplicated data can lead to overfitting problems: for this reason, it is appropriate to oversample the dataset with synthetic data.

Figure 5.1: Undersampling and oversampling




Over sampling by generating **synthetic data** is useful to increase the minority class to balance the dataset through the generation of new synthetic observations based upon the minority ones. There are different methods to oversample, in this way, an imbalanced dataset for a typical classification problem. One of the most common technique is called SMOTE (Synthetic Minority Over-sampling Technique), the one used in this research, and it looks at the feature space for the minority class data points and considers its k-nearest neighbors to create new synthetic points and increase the cardinality of the class itself (Fig. 5.2). The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the $k$ minority class nearest neighbors [73]. First it finds the nearest neighbors in the minority class for each of the samples in the class and then it consider a segment between the neighbors to generate random points on the segments. The SMOTE algorithm was applied on the real and complete records to create extra training data.

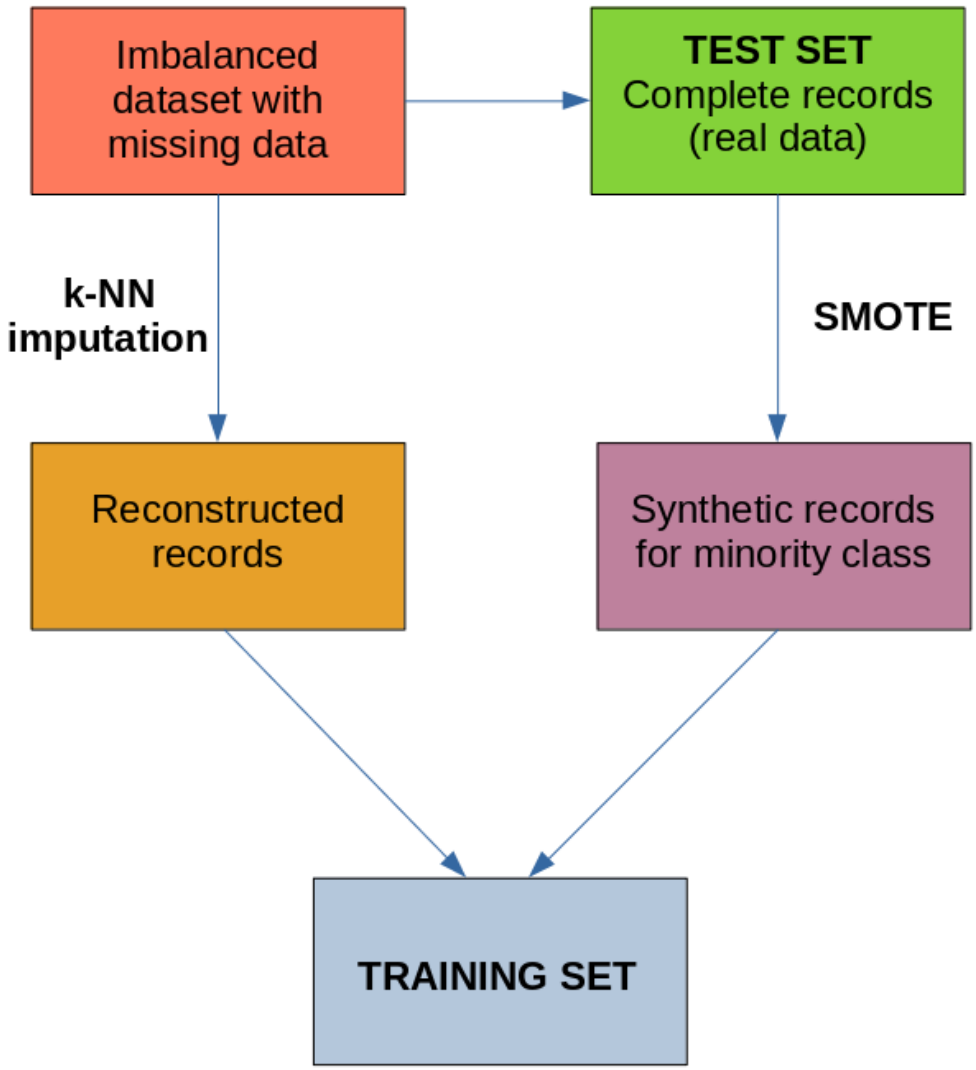
Figure 5.2: SMOTE visual representation




## 5.3 Composition and comparison of training and test set

After the data imputation and balancing stages we are able to compose a training set with enough data and balanced outcome. Figure 5.3 shows the main steps of the method for composing training and test set for the classification model. The training set is partly reconstructed and partly synthetic and the test set contains only the complete records or real data.

Figure 5.3: Training and test composition scheme




To verify the goodness of the training set records, we compare the distributions of each single variable with those of the test set. We observe if the training data are compliant with the real data. In the example in figures 5.4 and 5.5 we compare the distribution of a reconstructed continuous variable used in the training set with the real one used in the test set. The example in figure 5.6 shows the comparison between reconstructed and real binary variable, instead fig. 5.7 compare a categorical variable.

Figure 5.4: Distributions comparison

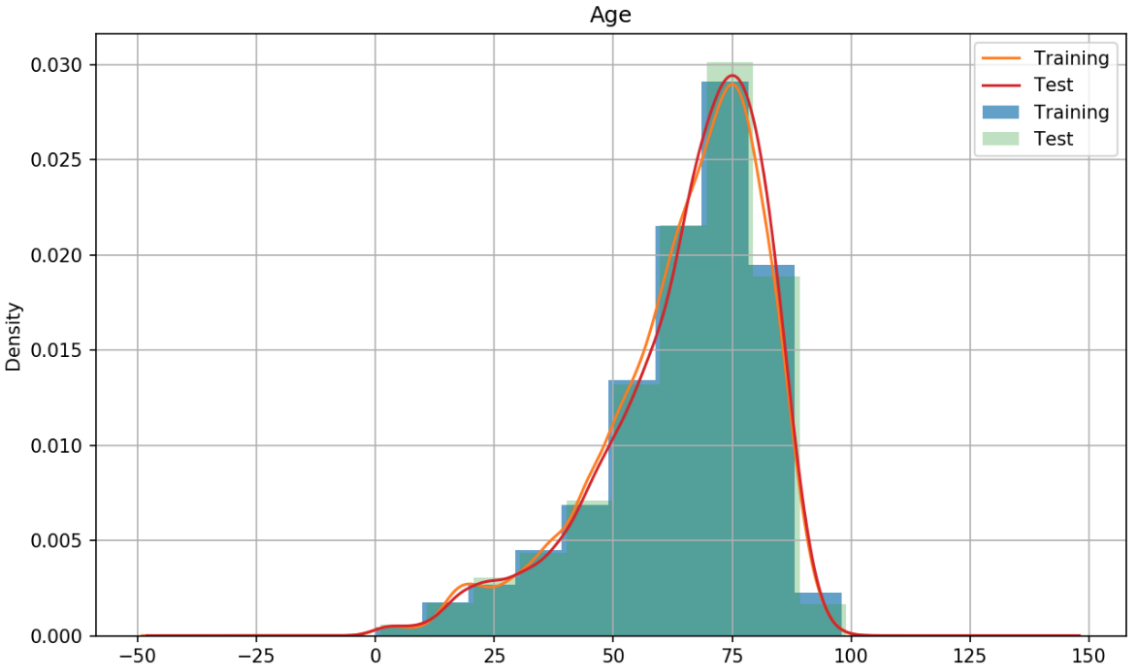


Figure 5.5: Continuous variables comparison

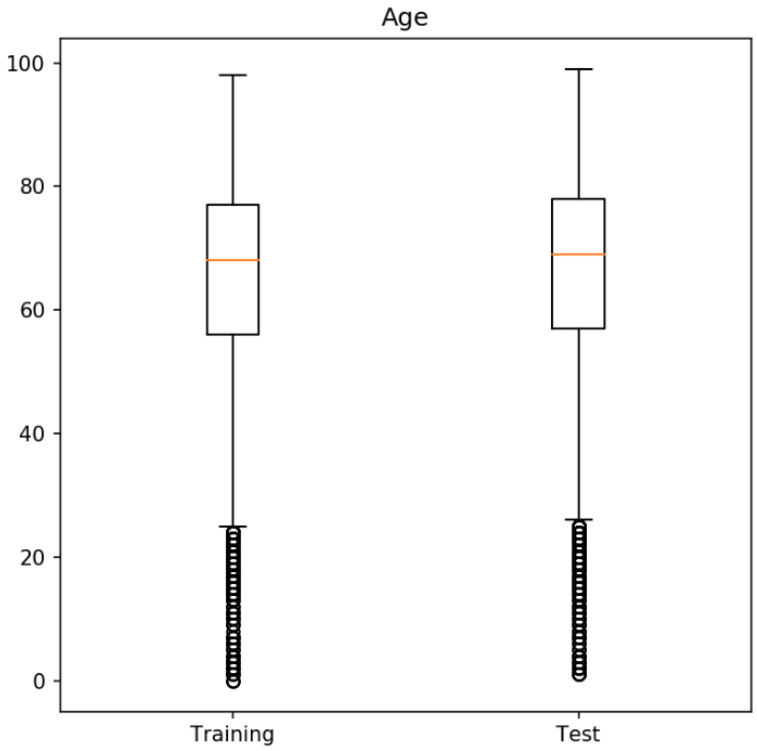



Figure 5.6: Dichotomous variables comparison

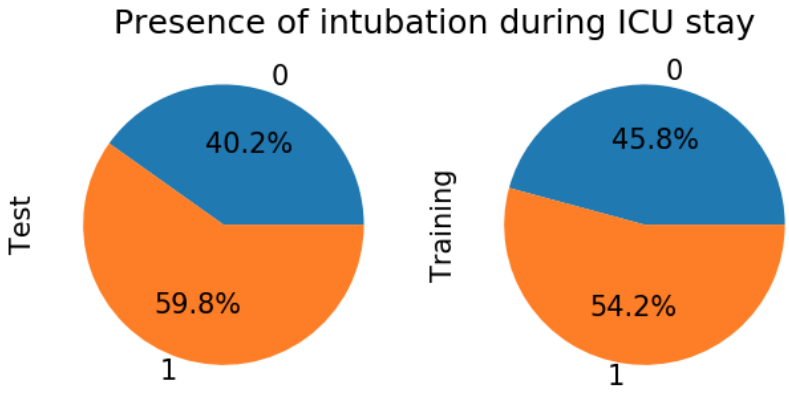
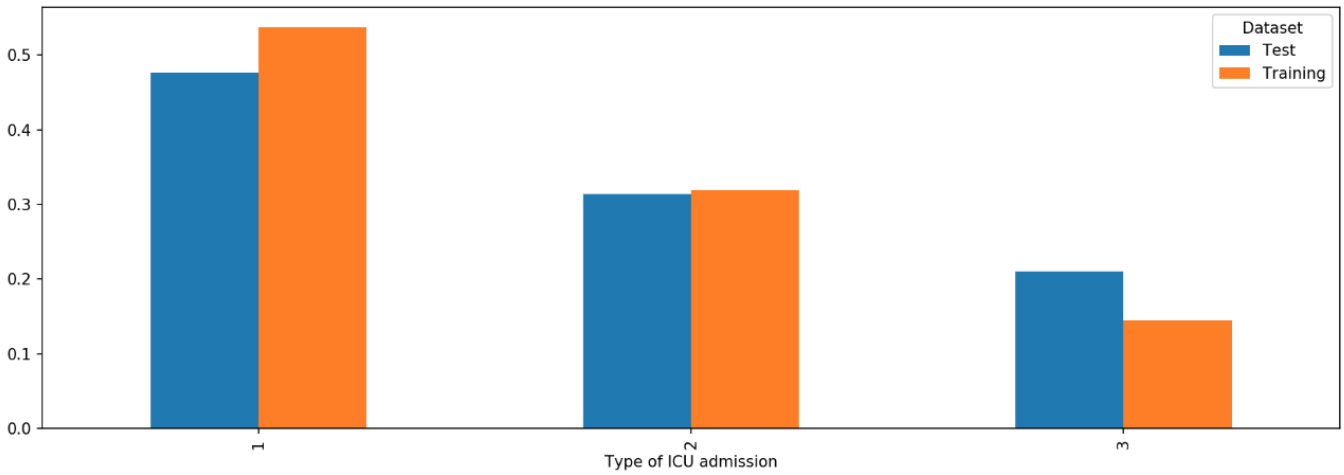


Figure 5.7: Categorical variables comparison




A further check to assess that the training data are compliant with the real data was performed by comparing the reconstructed and synthetic records with the real ones through the Andrews curve, considering 100 random observation (50 per class) per dataset in order to make the diagram more readable (eg. Fig. 5.8).

In data visualization, an Andrews plot or Andrews curve is a way to visualize structure in high-dimensional data. It is basically a smoothed version of a parallel coordinate plot.

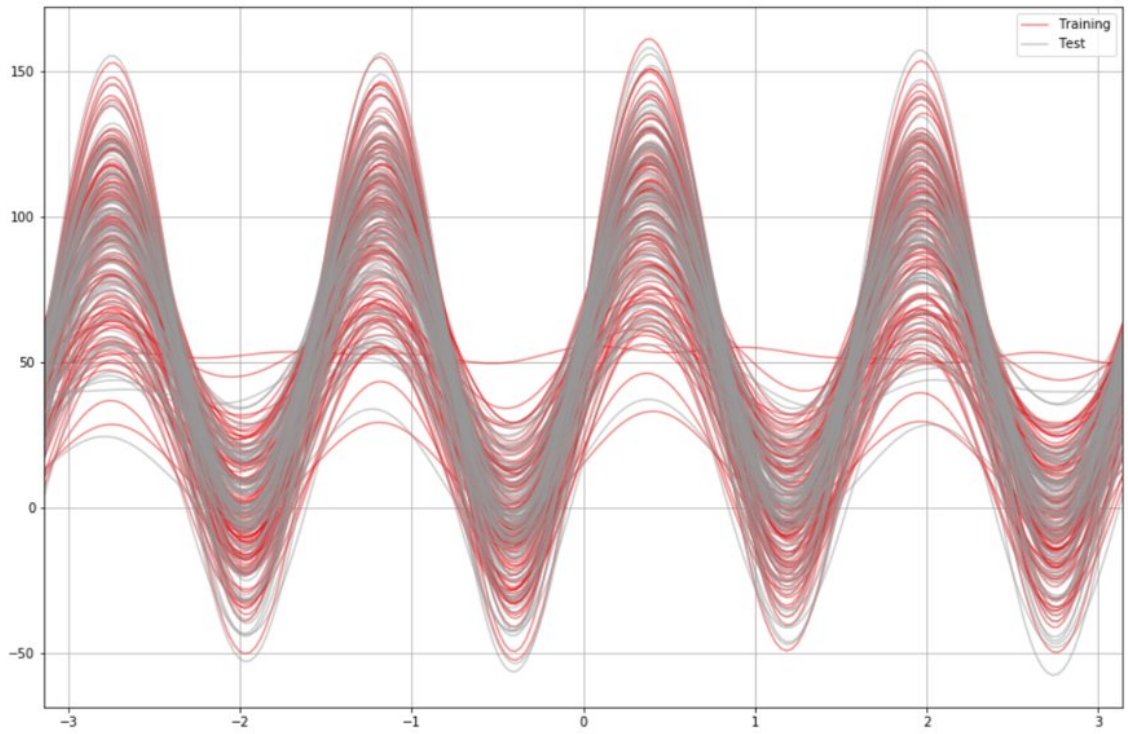
If we consider $d$ dimensions record $x = x_1, x_2, ..., x_d$, the Andrews plot defines a finite Fourier series

$$f_x(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \dots$$

and plot the curves for $-\pi < t < \pi$. In this way each data point may be represented as a line between $-\pi$ and $\pi$. This function can be seen as the projection of the data point onto the vector:

$$\left( \frac{1}{\sqrt{2}}, \sin(t), \cos(t), \sin(2t), \cos(2t), \dots \right)$$

Figure 5.8: Andrews plot of training and test samples




Once the compliance of the data obtained has been verified, we can proceed with the training of our classification model with the augmented data.

# Chapter 6

# Case studies

For some real-world applications, regressions and correlations are sufficient: if we want to know the general trend of a variable against another, for example, a simple correlation will determine the coefficient related to the trend, but if we want to classify non-linear separable data, Artificial Intelligence and Machine Learning can provide models and functions representing the closest possible match to the behavior of the data.

In the fields of bio-medicine and epidemiology, Machine Learning can be a powerful tool for finding correlations and patterns in data and it might represents a key component and a core element for Public Health policy-making. In the era of precision medicine, identifying patients at risk of HAIs by coupling established clinicalpathological features might be fundamental for developing novel preventive strategies tailored to each patients requirements [74, 75].

## 6.1 The SPIN-UTI network

In 2005, the Italian Study Group of Hospital Hygiene (GISIO) of the Italian Society of Hygiene, Preventive Medicine and Public Health (SItI) established the Italian Nosocomial Infections Surveillance in Intensive Care Units (SPIN-UTI) project [13, 76–81]. To date, the SPIN-UTI Network has surveyed approximately 20,000 patients, more than 4300 infections and 5300 micro-organisms [31] using the ECDC protocol [82].

Data has been initially stored in several different formats (SPSS spreadsheets, CSV, etc.). A preliminary work was therefore to clean, make uniform and merge the data of the different SPIN-UTI editions.

In general, the SPIN-UTI project prospectively surveys patients staying in ICU for more than two days and collects data at hospital, ICU and patient level. By contrast, patients who stay in ICU less than two days are excluded a priori. The reason for their exclusion is because the primary outcome of the SPIN-UTI project is the incidence of HAIs, which by definition develop after two days of ICU stay. The study was approved by the Ethics Committee Catania 1, Catania, Italy (protocol numbers 111/2018/PO and 295/2019/EMPO).

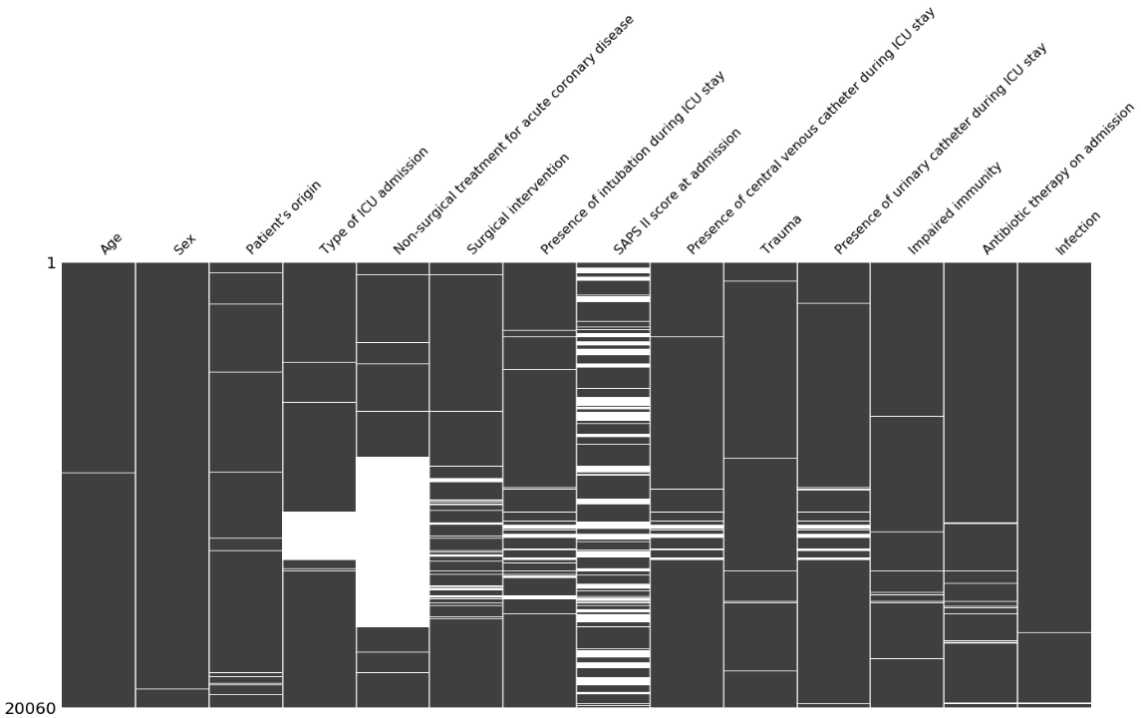
## 6.2 Models and subsets

### 6.2.1 Prediction of Healthcare-Associated Infections at ICU admission

#### Study design and data augmentation

For this case study the original dataset contained only 39% of patients (n=7827) with a complete assessment of variables considered in this study (figure 6.1).

Figure 6.1: Matrix of missing values



Since machine learning approaches require large data sets for training, we built a novel training data set made of recovered and synthetics data to tune the learning algorithms, together with a test set composed only by real data of patients with a complete assessment of the following variables at ICU admission: sex (dichotomous), patients origin (categorical: other ward/healthcare facility, community), non-surgical treatment for acute coronary disease (dichotomous), surgical intervention (dichotomous), SAPS II score at admission (continuous), presence of invasive devices at ICU admission (three dichotomous variables for urinary catheter, intubation and central venous catheter, respectively), trauma (dichotomous), impaired immunity (dichotomous), antibiotic therapy in 48 hours before ICU admission (dichotomous) [83].

Methods for data imputation, balancing and comparison were described in the previous chapter.

Starting from the most incomplete variable, with 2 cycles of 1-NN imputation applied separately to the two classes of data, infected patients or not, we have reconstructed 7758 records, approximately the 63% of the incomplete ones. After imputation, all available data was included in the analysis.

The dataset was, also, strongly imbalanced especially in terms of infected and not-infected classes. To avoid a low performance in recall score, on the 7827 complete record we applied SMOTE algorithm which provided a total of 2544 new synthetic records for the minority class (infected patients), also discarding the duplicated data generated by the algorithm.

### Training and Test Set composition and comparison

The training set is made by recovered ($N = 7758$) and synthetics records ($N = 2544$), while the test set includes 7827 real data. The distribution of infected (class 1) and non-infected (class 0) patients between the training and test sets is summarized in table 6.1. To evaluate the goodness of the training set, we compared the distributions of each single variable and the Andrews curves of the records with those of the test set to assess that the training data are compliant with the real data.

Table 6.1: Training and Test datasets composition

| | **Training Set** | **Test set** |
|---|---|---|
| **Class 0 (non-infected patients)** | 6702 reconstructed | 6602 |
| **Class 1 (infected patients)** | 3600 total<br>1056 reconstructed<br>2544 synthetics | 1225 |
| **Total** | **10302** | **7827** |

### Statistical Analysis

The Kolmogorov- Smirnov test was used to check the normal distribution of continuous variables. Patients characteristics were described using median and interquartile range (IQR) or percentage.

Comparisons between variables were analyzed by the Chi-squared test for categorical variables, while the Mann-Whitney U test was used for continuous variables with skewed distribution. To test the accuracy of the SAPS II score in HAIs risk prediction along the range of possible values, we used the Receiver Operating Characteristics (ROC) curve analysis. In particular, discrimination was assessed by calculating the area under the curve (AUC), with values ranging from 0.5 for no prediction to 1.0 for perfect prediction [11, 84, 85]. All statistical tests were two-sided, and $p - values < 0.05$ were considered statistically significant.

Statistical analyses were performed using SPSS software (version 26.0, SPSS, Chicago, IL)

### Learning model generation

To improve the predicting performance of the model, a machine learning algorithm combining the SAPS II with additional variables collected at ICU admission (i.e. sex, patients origin, non-surgical treatment for acute coronary disease, surgical intervention, presence of intubation, presence of urinary catheter, presence of central vascular catheter; trauma, impaired immunity, antibiotic therapy in 48 hours before ICU admission) was applied. Specifically, we chosen the SVM with Gaussian Kernel

(RBF) as modeling tool. This model has been successfully used in several regression and classification studies, especially for binary classification problems. Our model classifies data finding the best hyperplane separating the points of the classes. The separating hyperplane found by the algorithm provides the largest margin between the two classes. The selection of a non-linear kernel function, in our case the Gaussian kernel, is useful to map data that are not originally linearly separable into a higher dimensional feature space trying to make them more linearly separable. It is worth mentioning that linear kernels are less time consuming than non-linear ones, but they provides less accuracy.

Several datasets are not linearly separable even in a feature space, not allowing all of the constraints in the minimization problem of SVM to be satisfied [86]. To fill this gap, slack variables are introduced to allow certain constraints to be violated. By choosing very large slack variable values, we could find a degenerate solution which would lead to the model overfitting. To penalize the assignment of slack variables that were too large, the penalty is introduced in the classification objective:

$$C \sum_{i=1}^{N} \varepsilon_i$$

- $\varepsilon_i$, indicates slack variables, one for each datapoint $i$, to allow certain constraints to be violated;

- $C$, indicates a tuning parameter that controls the trade-off between the penalty of slack variables $\varepsilon_i$ and the optimization of the margin. High values of $C$ penalize slack variables leading to a hard margin, whereas low values of $C$ lead to a soft margin, which is a bigger corridor which allows certain training points inside at the expense of misclassifying some of them. In particular, the $C$ parameter sets the confidence interval range of the learning model.

The RBF kernel function expression on two sample, $x \; and \; x'$ , is defined as $K(x, x') = exp(-\gamma \|x - x'\|^2)$ where $\|x - x'\|^2$ is the squared Euclidean distance between the two feature vectors and $\gamma$ a is a free parameter. The RBF can be applied to a dataset through the choice of two parameters, $C$ and $\gamma$. The classifier performance of SVM depends on the choice of these two parameters. A grid search method was used to find the optimal parameters of RBF for SVM. This method

considered $m$ values in $C$ and $n$ values in $\gamma$, according to the $M \times N$ combination of $C$ and $\gamma$ [87] by training different SVM using a K-fold cross validation. Here, to optimize the *f1 score* of the positive class, we used a grid search on a 5-fold cross-validation.

Data analyses were performed through Python and the SciPy stack.

### Results

**Study population** On a total of 20060 SPIN-UTI participants, the current analysis was performed on a subsample of 7827 patients (median age = 69 years; 60.6% males) enrolled from 2006 to 2019. The remaining 12233 participants (61%) were excluded because of missing data on the assessment at ICU admission. In this subsample, patients coming from other wards/hospitals and reporting a surgical type of ICU admission were 73.9% and 52.4%, respectively. In general, median SAPS II score at admission was 40 (IQR = 28) and length of ICU stay was 5 days (IQR = 10). Patients who reported trauma and impaired immunity were 3.4% and 8.6%, respectively. With respect to medical treatments, 10.2% and 40.9% of patients underwent to non-surgical treatment for acute coronary disease or surgical intervention, while e 59% patients were on antibiotic therapy. In particular, the presence of urinary catheter, intubation and central venous catheter was 77.5%, 59.8% and 41%, respectively. Finally, we observed that percentage of ICU-acquired sepsis among patients enrolled was 6.1%, whereas ICU mortality was 23.2%.

**Characteristics of infected patients** Overall, table 6.2 also shows the comparison between infected ($N$ = 1225; 15.7%) and non-infected patients ($N$ = 6602; 84.3%) for characteristics at ICU admission. Infected patients were more likely to come from the community and to report a medical type of ICU admission than those non-infected. In particular, infected group consisted of patients who were more likely to report impaired immunity, also including more patients with trauma. This translated into higher SAPS II score among infected patients if compared with non-infected.

With respect to the presence of invasive devices, infected patients were also more likely to be intubated at ICU admission and less likely to be catheterized than those non-infected. As expected, infected patients exhibited higher length of ICU stay

(20.0 days vs. 4.0 days; $p < 0.001$) compared to non-infected patients. In line with these findings, also mortality was higher in infected patients (35.1%) than in those non-infected (21.0%; $p < 0.001$). No differences were evident for age, sex, non-surgical treatment for acute coronary disease, antibiotic therapy in 48 hours before ICU admission and presence of central venous catheter at ICU admission.

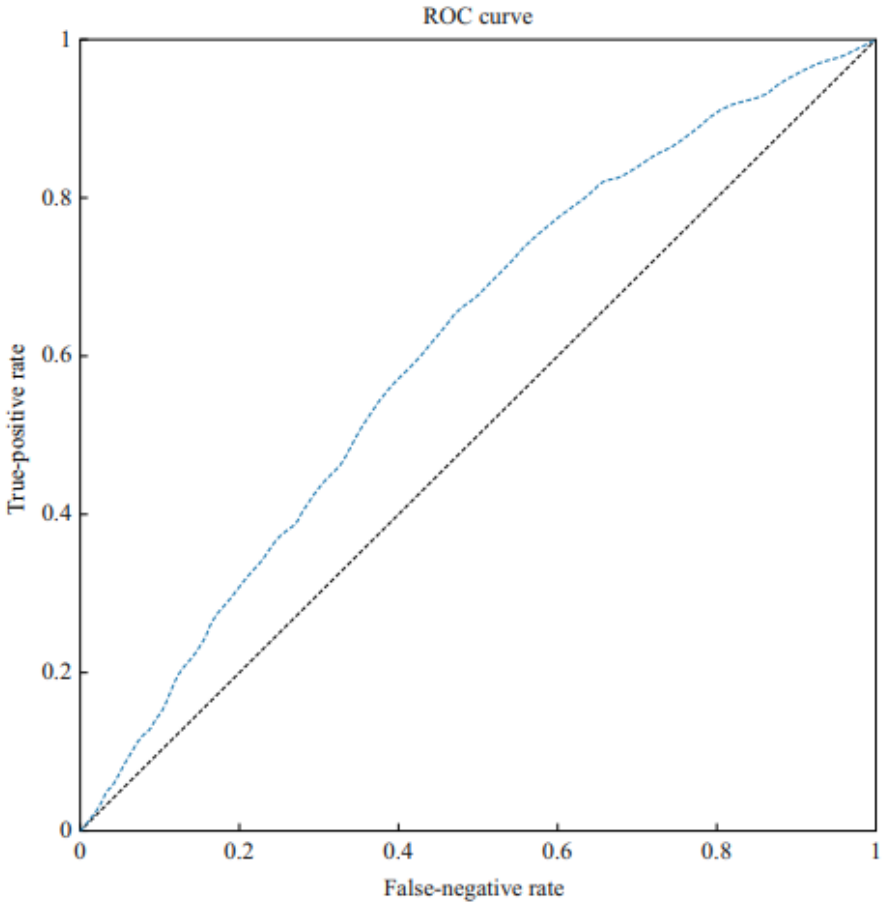
Table 6.2: Characteristics of patients according to their infectious status

| **Characteristics** | **Patients (N=7827)** | **Infected patients (N=1225)** | **Non-infected (N=6602)** | ***p-value*** |
|---|---|---|---|---|
| Age, years | 69.0 (21.0) | 69.0 (21.0) | 69.0 (21.0) | 0.064 |
| Sex (% male) | 60.6% | 62.8% | 60.1% | 0.084 |
| **Patient origin** | | | | |
| Other ward/healthcare facility | 73.9% | 67.7% | 75.1% | <0.001 |
| Community | 26.1% | 32.3% | 24.9% | |
| SAPS II at admission | 40.0 (28.0) | 47.0 (27.0) | 38.0 (27.0) | <0.001 |
| **Type of ICU admission** | | | | |
| Medical | 47.6% | 53.6% | 46.5% | <0.001 |
| Surgical | 52.4% | 46.4% | 53.5% | |
| Trauma | 3.4% | 5.0% | 3.2% | 0.001 |
| Impaired immunity | 8.6% | 10.4% | 8.2% | 0.015 |
| Non-surgical treatment for acute coronary disease | 10.2% | 8.9% | 10.4% | 0.109 |
| Surgical intervention | 40.9% | 36.7% | 41.7% | <0.001 |
| Antibiotic therapy in 48h preceding ICU admission | 59% | 59.8% | 58.9% | 0.579 |
| Presence of urinary catheter at ICU admission | 77.5% | 74.4% | 78.0% | 0.006 |
| Presence of intubation at ICU admission | 59.8% | 63.8% | 59.1% | 0.002 |
| Presence of central venous catheter at ICU admission | 41% | 39.7% | 41.3% | 0.295 |
| ICU-acquired sepsis (% yes) | 6.1% | 37.6% | - | - |
| Outcome (% death) | 23.2% | 35.1% | 21.0% | <0.001 |
| Length of ICU stay (days) | 5.0 (10.0) | 20.0 (20.0) | 4.0 (6.0) | <0.001 |

Results are reported as median (IQR) for continuous variables, or percentage for categorical variables. Statistical analyses were performed using MannWhitney test or Chi-squared test.

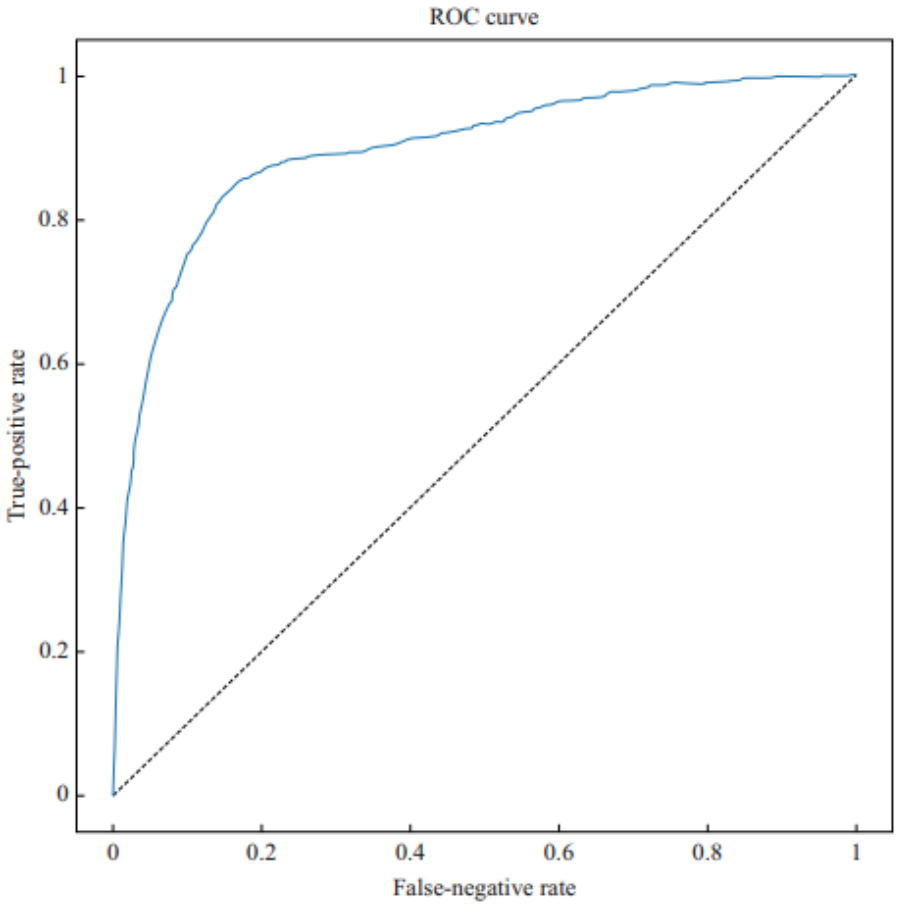
**ROC Curve Analysis using traditional statistical approach** Using traditional statistical analysis, we aimed to evaluate the performance of SAPS II score at ICU admission in predicting HAIs for all patients staying in ICU for more than two days. The Receiver operating characteristic (ROC) curve in figure 6.2 shows the ability of SAPS II to identify patients who developed at least one HAI during their stay in an intensive care unit. The curve plots the true-positive rate (i.e. sensitivity) vs the false-positive rate (i.e. 1 - specificity) at different classification thresholds. The blue curve represents the ability of SAPS II to distinguish between patients who developed at least one HAI and those who did not (AUC 0.612, 95% confidence interval $[0.60, 0.63]$; $p < 0.001$). The black dotted line is the reference for no predictive ability (AUC 0.500). Although this test was statistically significant, the accuracy of SAPS II score for predicting the risk of HAIs was of 56%.

Figure 6.2: ROC curve of the SAPS II for predicting HAIs



**ROC Curve Analysis using SVM model** To improve the accuracy for predicting the risk of HAIs, we employed the SVM algorithm, working on SAPS II score along with other characteristics at ICU admission. Figure 6.3 shows the ROC curve of SVM prediction model for the test set. We report that the accuracy of the SVM classifier was 88% on the test set. Specifically, precision and recall were 0.95 and 0.91 for non-infected patients and 0.60 and 0.73 for those who were diagnosed with at least one HAIs during their ICU stay. In line, the predictivity was assessed using ROC curve, which provided an AUC of 0.90 (95% Confidence Interval $= [0.88, 0.91]$; $p < 0.001$). Our results indicated the reliability of our SVM- model against overfitting.

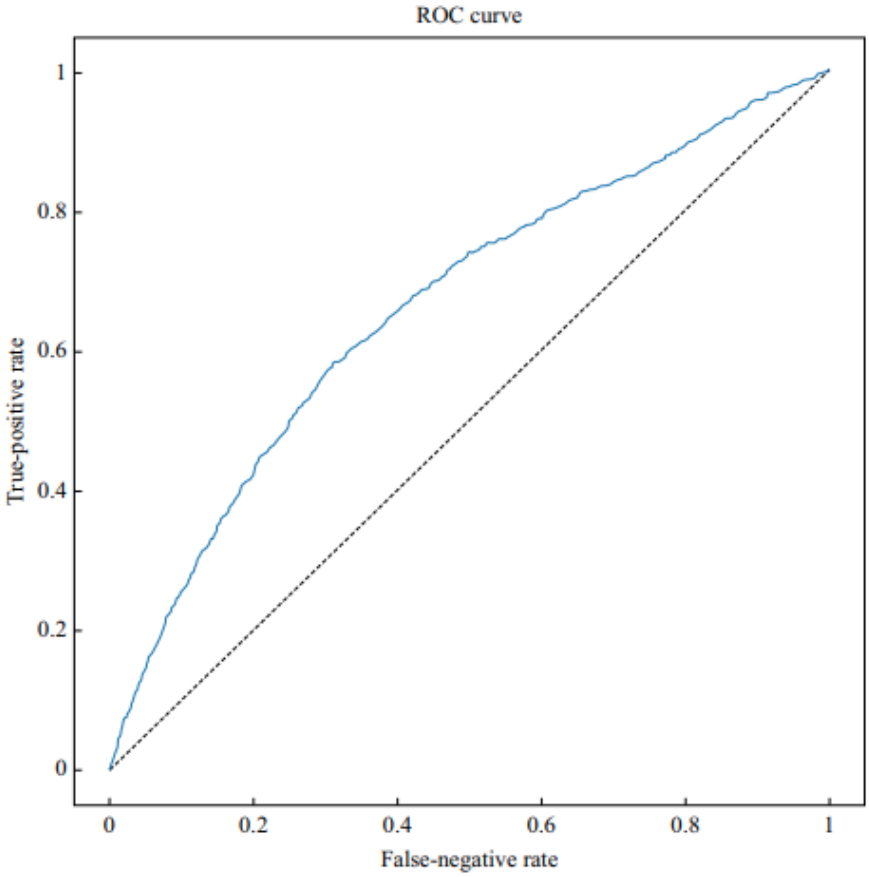
Figure 6.3: ROC curve of the SVM algorithm for predicting HAIs




Finally, we aimed to compare our prediction performance with those obtained on the same SVM model, without accounting for the SAPS II score variable in the

test set. Figure 6.4 shows the ROC curve of SVM prediction model for the test set, reporting an accuracy of 78%. Accordingly, precision and recall were 0.87 and 0.87 for non-infected patients and 0.31 and 0.32 for those infected, respectively. As expected, the AUC value provided by the ROC curve was 0.66 (95% Confidence Interval = $[0.65, 0.68]$; $p < 0.001$), indicating a lower predictive ability.

Figure 6.4: ROC curve of the SVM algorithm for predicting HAIs excluding the SAPS II




### Discussion

Identifying patients at higher risk of HAIs still represents a major challenge for public health, suggesting the need for novel tools that can guide patient management in ICUs [88–90]. Machine learning systems have been developed in many fields of medicine including infectious diseases control and clinical decision support [91]. Particularly, machine learning technique has been applied in patients with sepsis [92], to predict candidemia [93] or complications related to *Clostridium difficile* infection

[94], to improve the prediction of antimicrobial resistance [95], and for surveillance purpose [96].

To the best of our knowledge, the present study is the first one employing machine learning methods to identify patients at higher risk of HAIs, according to their individual characteristics at ICU admission. Indeed, there is current consensus that machine learning algorithms could support and enrich conventional statistical approaches, especially in terms of prediction of ICU prognosis, clinical deterioration and risk assessment [97–99]. Several modifiable and non-modifiable risk factors might affect the risk of HAIs and related adverse outcomes [100]. For instance, the prolonged use of invasive devices, patients impaired immunity, surgical intervention and comorbidity represent the major risk factors for HAIs in ICU [100, 101].

In clinical practice, several prognostic scores are routinely used to evaluate the complex clinical-pathological conditions of ICU patients, in order to develop novel and more suitable preventive strategies tailored to each patients requirements [102], [75]. For instance, the SAPS II score represents the most useful tool for the prediction of prognosis, HAIs risk, sepsis and mortality 9-11, 13, [11–13, 84, 102].

To this aim, we first evaluated the ability of SAPS II score at ICU admission for predicting HAIs risk of 7827 patients staying in ICU for more than two days. Interestingly, our ROC curve analysis, which provides an AUC value of 0.612, does not suggest the use of SAPS II score in the end-of-life decision-making. Indeed, although the test was statistically significant, the accuracy of SAPS II score for predicting the risk of HAIs was of 56%.

In this scenario, machine learning approaches represent a possible strategy for healthcare facilities, making possible to build a specific prediction model targeted to demographics and clinical characteristics of patients [97, 98]. In line, several studies suggested SVM technique as being an excellent and powerful algorithm to predict common complex diseases with many risk factors, having a better discrimination than conventional statistical approaches [22].

Accordingly, we employed the SVM algorithm, considering SAPS II score along with other characteristics at ICU admission (i.e. age, sex, SAPS II score at admission, patients origin, type of admission, trauma, impaired immunity, non-surgical

treatment for acute coronary disease, surgical intervention, presence of invasive devices, and antibiotic therapy), in order to improve the accuracy for predicting the risk of HAIs. Our findings demonstrated that the accuracy of the SVM classifier was 88% on the test set, reporting precision and recall values of 0.95 and 0.91 for non-infected patients and 0.60 and 0.73 for those who were diagnosed with at least one HAIs during their ICU stay. In line, the predictive ability assessed by the ROC curve provided an AUC of 0.90.

To assess the relevance of patients characteristic at ICU admission in our SVM model, we compared the prediction performance with those obtained by same SVM model, without accounting for the SAPS II score. We found a ROC curve reporting an accuracy of 78%, with precision and recall values of 0.87 and 0.87 for non-infected patients and 0.31 and 0.32 for those infected, respectively. Notably, the AUC value provided by the ROC curve was of 0.66, indicating a lower predictive ability. Due to its low predictive ability, our findings not warrant clinical usefulness of SAPS II score when considered alone, suggesting the need of an integrated approach with patients personal and clinical characteristics, which are crucial in determining the risk of HAIs and adverse outcomes in ICU.

Our findings provide a promising evaluation of a better predictive performance of the SVM algorithm than conventional statistical approaches, suggesting the SVM as a possible tool to identify and predict patients at higher risk of HAIs at ICU admission, providing clinicians sufficient time to potentially prevent HAI and mitigate its severity, targeting specific infection prevention and control interventions to high-risk groups in order to improve quality of care. Although further efforts are needed, predictive models in healthcare systems represent a useful strategy for better diagnosis, prognosis and personalized patients management, including preventive strategies against HAIs.

### 6.2.2 Early prediction of 7-days mortality in ICU using a machine learning model

In the present study, we aimed to identify and predict patients at higher risk of dying, considering their clinical and pathological characteristics at ICU admission using the data collected during the seven editions of the SPIN-UTI project. The

primary purpose of this study is to evaluate the ability of the SAPS II to predict the risk of death after 7 days from their admission to ICU. The secondary purpose is to develop and test a machine learning algorithm, which combines the SAPS II with additional patients characteristics, to further improve the predicting performance.

The data augmentation methods used previously were applied in this case study for a specific subset of the SPIN-UTI dataset.

### Definition of SAPS II and other predictors

SAPS II at ICU admission was initially used as the main predictor and its computation included the following components: age; heart rate; systolic blood pressure; temperature; Glasgow Coma Scale; continuous positive airway pressure; PaO2; FiO2; urine output; blood urea nitrogen; sodium; potassium; bicarbonate; bilirubin; white Blood Cell; chronic diseases; type of admission. Each component was assessed within 24 hours from ICU admission and the worst value was recorded. The total SAPS II was finally computed as the sum of weighted values for each component [19]. The SPIN-UTI project also collected information on patients who underwent non-surgical treatment for signs and symptoms related to the acute coronary syndrome. Moreover, we defined admission with trauma those resulting from blunt or penetrating traumatic injury to the patient, with or without surgical intervention. Instead, impaired immunity was defined as an impairment due to treatment, diseases or $< 500 PMN/mm^3$. Finally, we also collected if any antibiotic therapy was administered in the 48 hours preceding ICU admission and/or during the first two days of ICU stay. The occurrence of HAI was defined according to a set of clinical and laboratory criteria that are fully described in the ECDC protocol [103].

### Dataset of real records

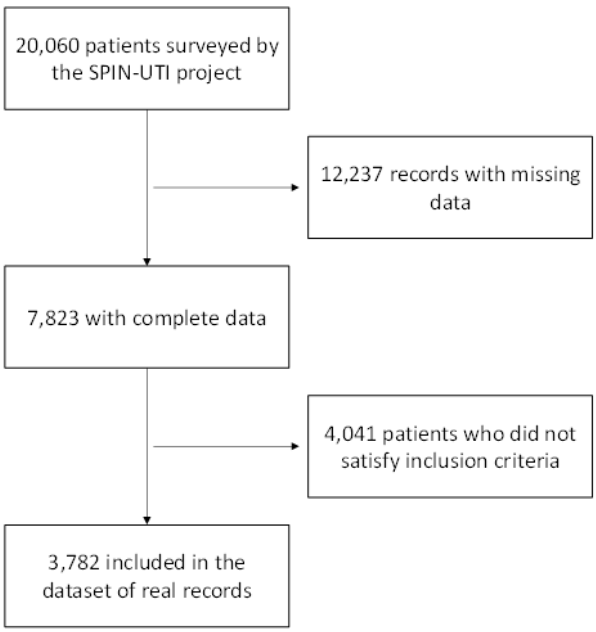
We first worked only on patients with a complete assessment of the following information: sex, patients origin, type of ICU admission, non-surgical treatment for acute coronary disease, surgical intervention, SAPS II, presence of invasive devices at ICU admission, trauma, impaired immunity, antibiotic therapy in 48 hours before or after ICU admission and onset of HAI. The primary outcome of the current analysis was mortality within seven days from ICU admission. Accordingly, the current analysis included:

1. patients who stayed in ICU for at least seven days;

2. those who died within two to seven days after ICU admission.

By contrast, patients who stayed in ICU for less than two days and those who died with the first two days were excluded. Figure 6.5 shows the scheme of the selection of real records. Specifically, this dataset of real records consisted on a total of 3782 patients with complete data and meeting the inclusion criteria. The dataset described above was used for traditional statistical analyses and as test set for the machine learning algorithm.

Figure 6.5: Selection of records with complete data satisfying inclusion criteria




### Dataset of synthetic records

In the current subset of the SPIN-UTI dataset there were 61% ($N = 12237$) of records with missing data. As many statistical models and machine learning algorithms rely on complete datasets, it is key to handle the missing data appropriately. Moreover, machine learning algorithms generally requires large datasets to be trained. For these reasons, we created a dataset of synthetic records that was used as training set for the machine learning algorithm. Accordingly, we first imputed

missing data from incomplete records of the original dataset, using the K-Nearest Neighbor (K-NN) imputation method.

Two cycles of K-NN imputation to the two classes of data (i.e. alive or died patients) reconstructed 3258 records that satisfied inclusion criteria [29]. Moreover, synthetic data were generated to balance the two classes of died and alive patients, using the Synthetic Minority Over-sampling Technique (SMOTE).

The SMOTE alghoritm on 3782 real records, obtained 1131 synthetic records for the class of died patients. Given that, the dataset of synthetic records, which was used as the training set, included a total of 4389 records (table 6.3). To confirm the goodness of the training set, we compared the distributions of primary outcome and exposure variables with those obtained from the test set (figs. 6.6 to 6.9) and the Andrews plot of the observations (fig. 5.8).

Methods used for data augmentation are reported in the previous chapter.

Table 6.3: Training set and Test set composition

| | **Training Set** | **Test set** |
|---|---|---|
| **Class 0 (alive patients)** | 2596 recovered | 2907 |
| **Class 1 (dead patients)** | 1193 total<br>662 reconstructed<br>1131 synthetics | 875 |
| **Total** | **4589** | **3782** |

Figure 6.6: Age distribution of training and test set

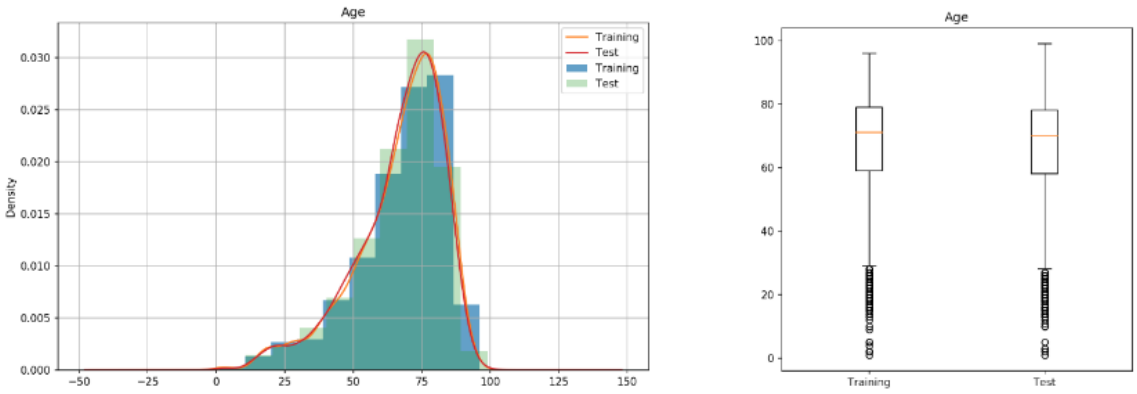



Figure 6.7: SAPS II distribution of training and test set

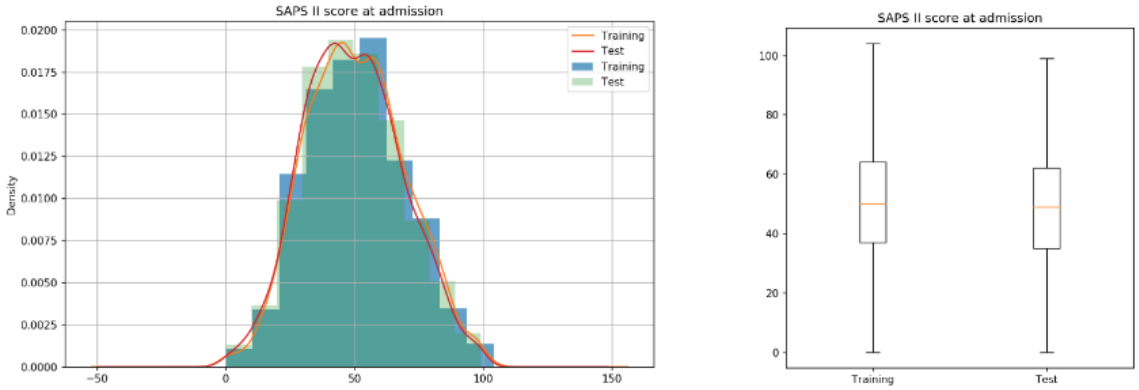


Figure 6.8: Binary variables comparison

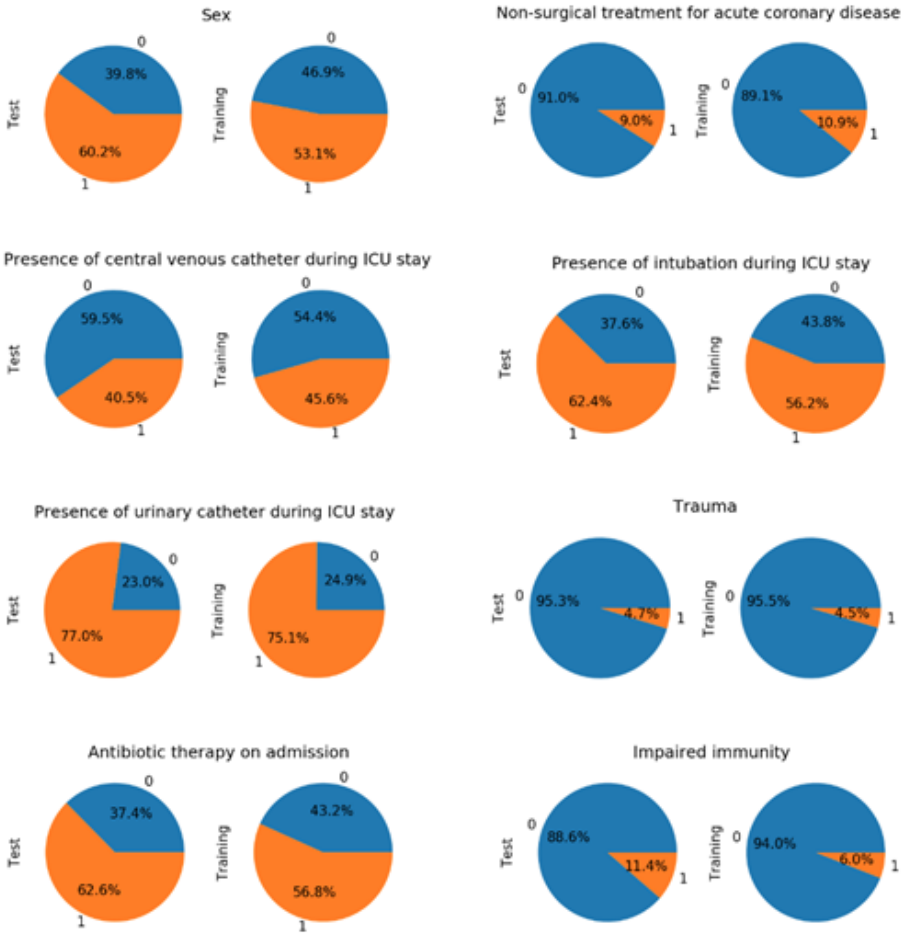
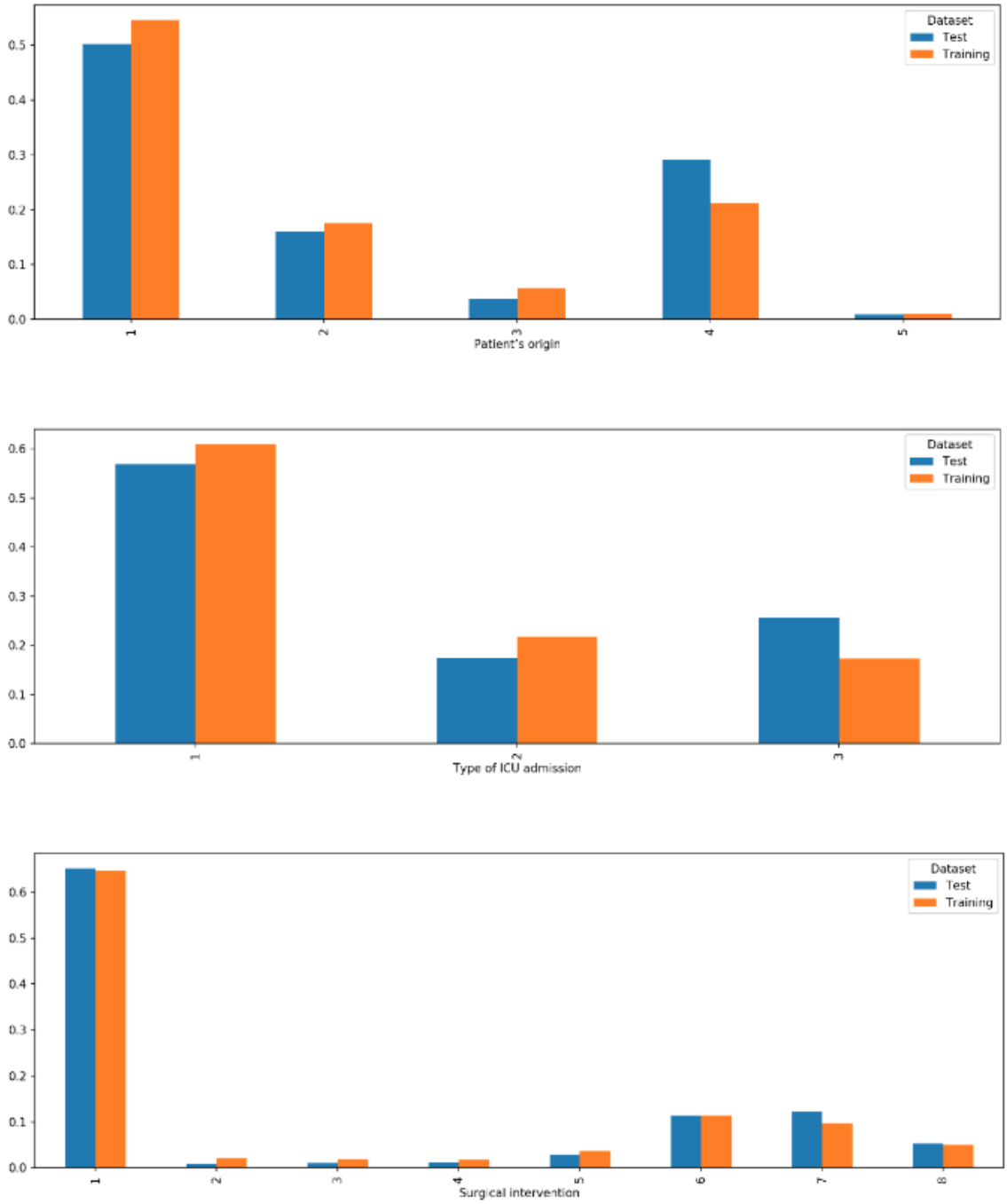


Figure 6.9: Categorical variables of training and test set



### Statistical analysis

All variables of the real dataset were described according to their type and skewness using descriptive statistics (frequencies and percentages [%] or median and interquartile range [IQR]). In an epidemiological and descriptive point of view, we compared these variables between dead and alive patients using the Mann-Whitney U test for quantitative variables and the Chi-Squared test and Chi-Squared for trend test for qualitative variables. We first used a logistic regression model to evaluate the association of SAPS II (continuous) with death. Next, we applied a logistic regression model, also including sex (dichotomous), patients origin (categorical: other ward/healthcare facility, community), type of ICU admission (categorical: medical, surgical), non-surgical treatment for acute coronary disease (dichotomous), surgical intervention (dichotomous), presence of invasive devices at ICU admission (three dichotomous variables for urinary catheter, intubation and central venous catheter, respectively), trauma (dichotomous), impaired immunity (dichotomous), antibiotic therapy in 48 hours before or after ICU admission (dichotomous). We also used Receiver Operating Characteristic (ROC) curves to assess the ability of the logistic regression models to accurately identify patients who dead from those who did not. Results were reported in terms of Area Under the Curve (AUC) and 95% Confidence Interval (95%CI). With respect to the model on SAPS II alone, we identified the best cut-off value which maximized the Youden Index. For the best cut-off value, sensitivity and specificity with their 95%CI were calculated. All tests will be performed at a significance level $\alpha = 0.05$ and statistical analysis will be conducted using SPSS v.25.

### Machine Learning Algorithm

We next compared the predictive performance of 7-day mortality between logistic regression model and a machine learning algorithm. Specifically, the algorithm combined SAPS II with the following variables collected at ICU admission: sex, patients origin, type of ICU admission, non-surgical treatment for acute coronary disease, surgical intervention, presence of intubation, presence of urinary catheter, presence of central vascular catheter; trauma, impaired immunity, antibiotic therapy in 48 hours before or after ICU admission. For the current analysis, we chosen the supervised SVM algorithm as modelling tool, which can be used for classification -

especially for binary classification - and regression problems. However, our dataset was not linearly separable, not allowing to satisfy all the constraints of SVM. For this reason, we used a non-linear Kernel function (i.e., the Gaussian Kernel, also called as Radial basis function Kernel, RBF). Slack variables with penalty were also introduced to satisfy all the constraints in the minimization problem of SVM [86]. The SVM model was trained on the training set composed by synthetic records, and then tested on the test set made of real records. Since patients who developed HAIs during their ICU stay are generally at higher risk of death, we also tested the SVM model on those who did not acquire HAIs within seven days from ICU admission. We also assessed the predictive performance of a SVM model, which included all variables collected at ICU admission except of SAPS II. Results are reported in terms of AUC, accuracy, sensitivity, and specificity with their 95%CI. The analyses were performed using Python and Support Vector Classification (SVC) from Sklearn 0.22.1.

### Results

**Characteristics of the dataset of real records** The current analysis included 3,782 SPIN-UTI participants without missing data (60.2% males), surveyed from 2006 to 2019. In this subsample, the median age was 70.0 years (IQR=20) and median SAPS II score at admission was 49 (IQR=27). Overall, 70.9% came from other wards/hospitals and 56.9%, reported a medical type of ICU admission. In particular, 4.7% and 11.4% of patients reported trauma and/or impaired immunity, respectively. Patients who underwent antibiotic therapy, surgical intervention or non-surgical treatment for acute coronary disease were 62.6%, 34.8% and 9.0%, respectively. With respect to invasive devices, the presence of urinary catheter, intubation and central venous catheter was reported in 77.0%, 62.4% and 40.5% patients, respectively. Table 6.4 compares characteristics of patients who died ($N$ = 875; 23.1%) within seven day from ICU admission with those who were still alive ($N$ = 2907; 76.9%). Specifically, patients who died were older, more likely men, and with a higher SAPS II than those who did not die. Moreover, they were also more likely to come from other ward/ healthcare facility and to report a medical type of ICU admission than those alive. The first group also consisted more of patients who reported impaired immunity and less traumatic events. Instead, no differences were

evident for surgical intervention, non-surgical treatment for acute coronary disease, antibiotic therapy on admission and presence of invasive devices at ICU admission.

Table 6.4: Characteristics of patients with complete data according to their outcome status

| **Characteristics** | **Patients (N=3782)** | **Dead patients (N=875)** | **Alive patients (N=2907)** | ***p-value*** |
|---|---|---|---|---|
| Age, years | 70.0 (20.0) | 74.0 (17.0) | 69.0 (21.0) | <0.001 |
| Sex (% male) | 60.2% | 55.0% | 61.7% | <0.001 |
| **Patient origin** | | | | |
| Other ward/healthcare facility | 70.9% | 70.1% | 71.1% | <0.001 |
| Community | 29.1% | 29.9% | 28.9% | |
| SAPS II at admission | 49.0 (27.0) | 59.0 (27.0) | 46.0 (25.0) | <0.001 |
| **Type of ICU admission** | | | | |
| Medical | 56.9% | 59.3% | 56.1% | <0.001 |
| Surgical | 43.1% | 40.7% | 43.9% | |
| Trauma | 4.7% | 2.4% | 5.4% | <0.001 |
| Impaired immunity | 11.4% | 15.0% | 10.3% | <0.001 |
| Non-surgical treatment for acute coronary disease | 9.0% | 10.2% | 8.7% | 0.174 |
| Surgical intervention | 34.8% | 32.5% | 35.5% | 0.306 |
| Antibiotic therapy in 48h preceding ICU admission | 62.6% | 62.2% | 62.8% | 0.744 |
| Presence of urinary catheter at ICU admission | 77.0% | 75.9% | 77.4% | 0.351 |
| Presence of intubation at ICU admission | 62.4% | 61.7% | 62.6% | 0.646 |
| Presence of central venous catheter at ICU admission | 40.5% | 38.5% | 41.0% | 0.182 |

Results are reported as median (IQR) for continuous variables, or percentage for categorical variables. Statistical analyses were performed using MannWhitney test or Chi-squared test.

### Applying logistic regression models to predict the risk of 7-day mortality

We first applied a logistic regression model on the dataset of real records, using SAPS II as the independent and 7-day mortality as the dependent variable. Accordingly, Figure 6.10 illustrates the accuracy of SAPS II for predicting the risk of 7-day mortality for all patients admitted in ICU. We noted that SAPS II was able to

discriminate patients who died from those who did not, with AUC of 0.678 (95%CI = $[0.657, 0.700]$; $p < 0.001$) and accuracy of 69.3% (95%CI = $[67.8\%, 70.8\%]$). The coordinates of the ROC curve are reported in table 6.5. Specifically, the best cut-off value of SAPS II, which maximized the Youden index, was 54.5. The application of this value resulted in sensitivity of 61.9% (95%CI = $[60.4\%, 63.4\%]$) and specificity of 67.1% (95%CI = $[65.6\%, 68.7\%]$). We further applied a logistic regression model, which combined SAPS II with additional patients characteristics collected at ICU admission. However, as indicated in figure 6.11, both AUC and accuracy of this model remained moderate (AUC = 0.637; 95%CI = $[0.616, 0.659]$; Accuracy = 65.2%; 95%CI = $[63.7\%, 66.7\%]$). In line, sensitivity (true positive rate) and specificity (true negative rate) for death were 49.0% (95%CI = $[47.5\%, 50.5\%]$) and 70.0% (95%CI = $[68.5\%, 71.5\%]$), respectively.

Figure 6.10: ROC curves of logistic regression models to predict 7-day mortality using SAPS alone

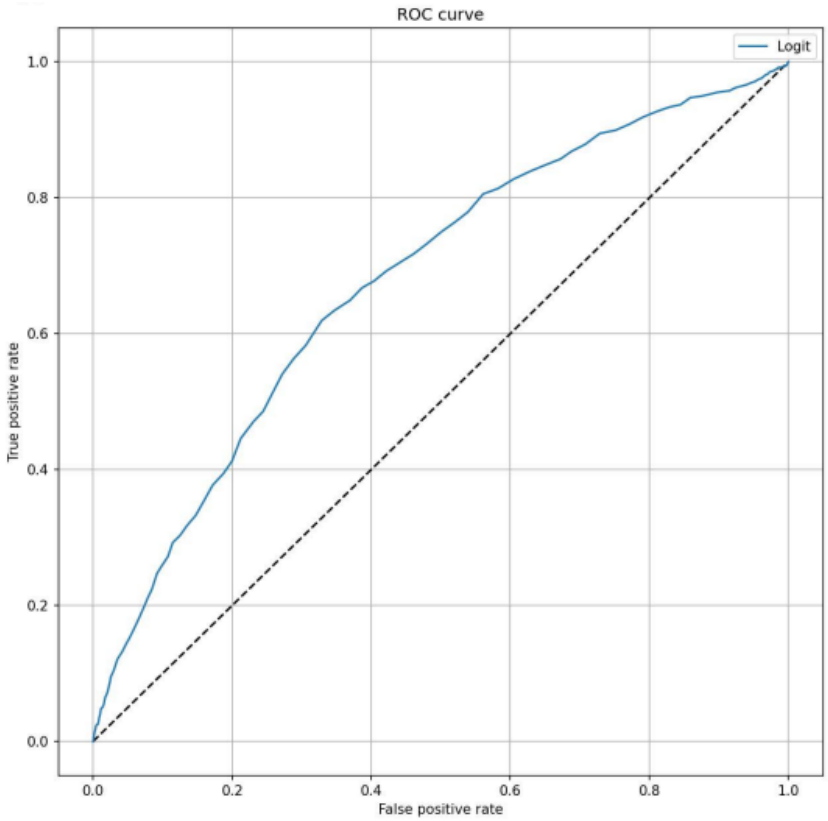
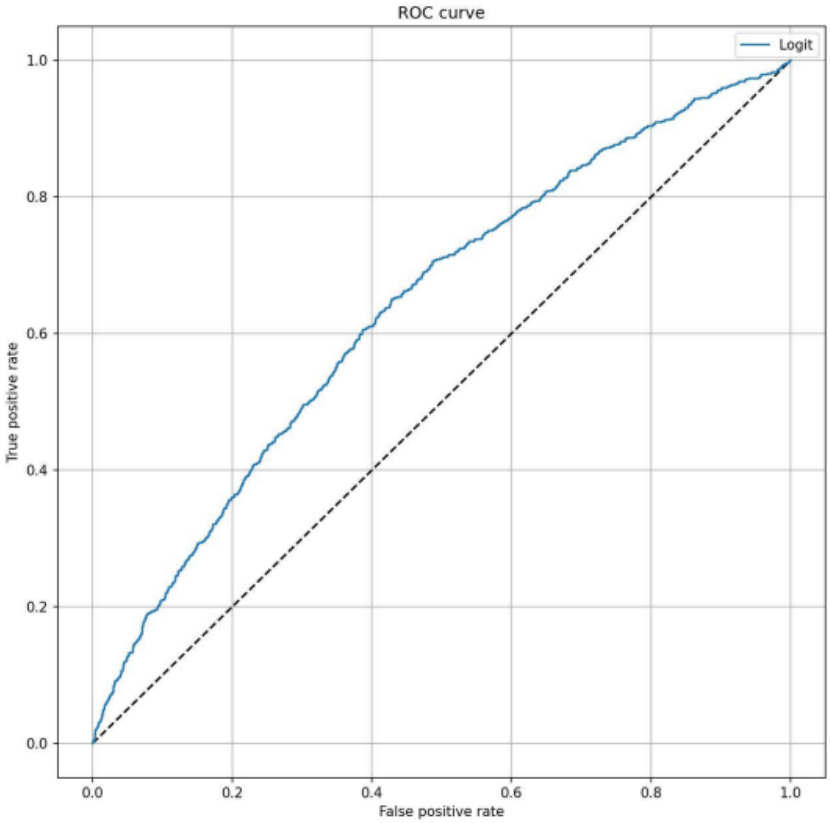


Table 6.5: Coordinates of the ROC curve of logistic regression model with SAPS II alone

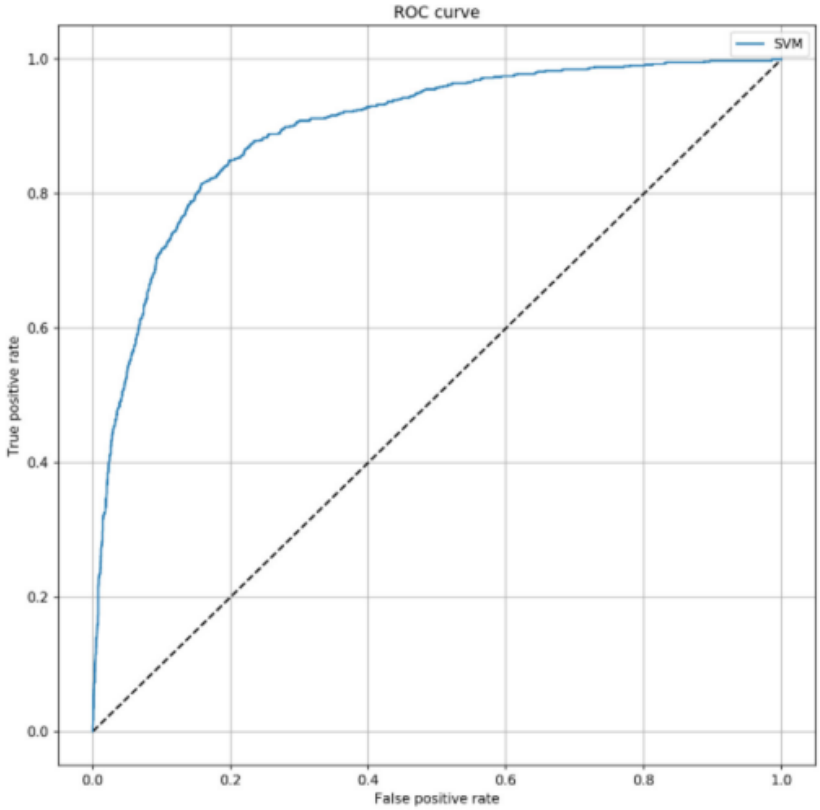
| **SAPS II values** | **TPR** | **FPR** | **SAPS II values** | **TPR** | **FPR** | **SAPS II values** | **TPR** | **FPR** |
|---|---|---|---|---|---|---|---|---|
| **1** | 0.997 | 0.998 | **34** | 0.899 | 0.751 | **67** | 0.333 | 0.148 |
| **2** | 0.995 | 0.998 | **35** | 0.895 | 0.729 | **68** | 0.318 | 0.135 |
| **3** | 0.995 | 0.997 | **36** | 0.879 | 0.708 | **69** | 0.303 | 0.125 |
| **4** | 0.994 | 0.997 | **37** | 0.869 | 0.688 | **70** | 0.293 | 0.115 |
| **5** | 0.994 | 0.996 | **38** | 0.857 | 0.672 | **71** | 0.272 | 0.108 |
| **6** | 0.993 | 0.994 | **39** | 0.848 | 0.649 | **72** | 0.258 | 0.099 |
| **7** | 0.993 | 0.992 | **40** | 0.839 | 0.628 | **73** | 0.247 | 0.092 |
| **8** | 0.992 | 0.990 | **41** | 0.827 | 0.604 | **74** | 0.224 | 0.085 |
| **9** | 0.992 | 0.987 | **42** | 0.814 | 0.582 | **75** | 0.211 | 0.079 |
| **10** | 0.992 | 0.986 | **43** | 0.806 | 0.561 | **76** | 0.197 | 0.073 |
| **11** | 0.990 | 0.984 | **44** | 0.779 | 0.539 | **77** | 0.178 | 0.065 |
| **12** | 0.990 | 0.982 | **45** | 0.765 | 0.521 | **78** | 0.161 | 0.057 |
| **13** | 0.987 | 0.980 | **46** | 0.750 | 0.501 | **79** | 0.152 | 0.052 |
| **14** | 0.985 | 0.973 | **47** | 0.731 | 0.479 | **80** | 0.144 | 0.048 |
| **15** | 0.983 | 0.971 | **48** | 0.718 | 0.461 | **81** | 0.131 | 0.042 |
| **16** | 0.981 | 0.967 | **49** | 0.704 | 0.440 | **82** | 0.121 | 0.035 |
| **17** | 0.976 | 0.961 | **50** | 0.693 | 0.422 | **83** | 0.104 | 0.030 |
| **18** | 0.975 | 0.958 | **51** | 0.678 | 0.404 | **84** | 0.095 | 0.025 |
| **19** | 0.971 | 0.952 | **52** | 0.667 | 0.387 | **85** | 0.087 | 0.024 |
| **20** | 0.969 | 0.946 | **53** | 0.649 | 0.369 | **86** | 0.072 | 0.021 |
| **21** | 0.967 | 0.941 | **54** | 0.634 | 0.347 | **87** | 0.064 | 0.017 |
| **22** | 0.965 | 0.933 | **55** | 0.619 | 0.329 | **88** | 0.055 | 0.015 |
| **23** | 0.962 | 0.924 | **56** | 0.583 | 0.306 | **89** | 0.050 | 0.014 |
| **24** | 0.958 | 0.915 | **57** | 0.563 | 0.289 | **90** | 0.048 | 0.011 |
| **25** | 0.955 | 0.899 | **58** | 0.541 | 0.272 | **91** | 0.039 | 0.010 |
| **26** | 0.953 | 0.890 | **59** | 0.512 | 0.258 | **92** | 0.035 | 0.009 |
| **27** | 0.950 | 0.875 | **60** | 0.486 | 0.245 | **93** | 0.031 | 0.008 |
| **28** | 0.947 | 0.859 | **61** | 0.471 | 0.231 | **94** | 0.025 | 0.007 |
| **29** | 0.937 | 0.845 | **62** | 0.446 | 0.212 | **95** | 0.025 | 0.006 |
| **30** | 0.934 | 0.829 | **63** | 0.413 | 0.200 | **96** | 0.023 | 0.004 |
| **31** | 0.927 | 0.810 | **64** | 0.394 | 0.187 | **97** | 0.018 | 0.003 |
| **32** | 0.919 | 0.791 | **65** | 0.377 | 0.172 | **98** | 0.016 | 0.003 |
| **33** | 0.909 | 0.772 | **66** | 0.353 | 0.159 | **99** | 0.011 | 0.002 |

Figure 6.11: ROC curves of logistic regression models to predict 7-day mortality including sex, patients origin, type of ICU admission, non-surgical treatment for acute coronary disease, surgical intervention, presence of invasive devices at ICU admission, trauma, impaired immunity, antibiotic therapy in 48 hours before or after ICU admission



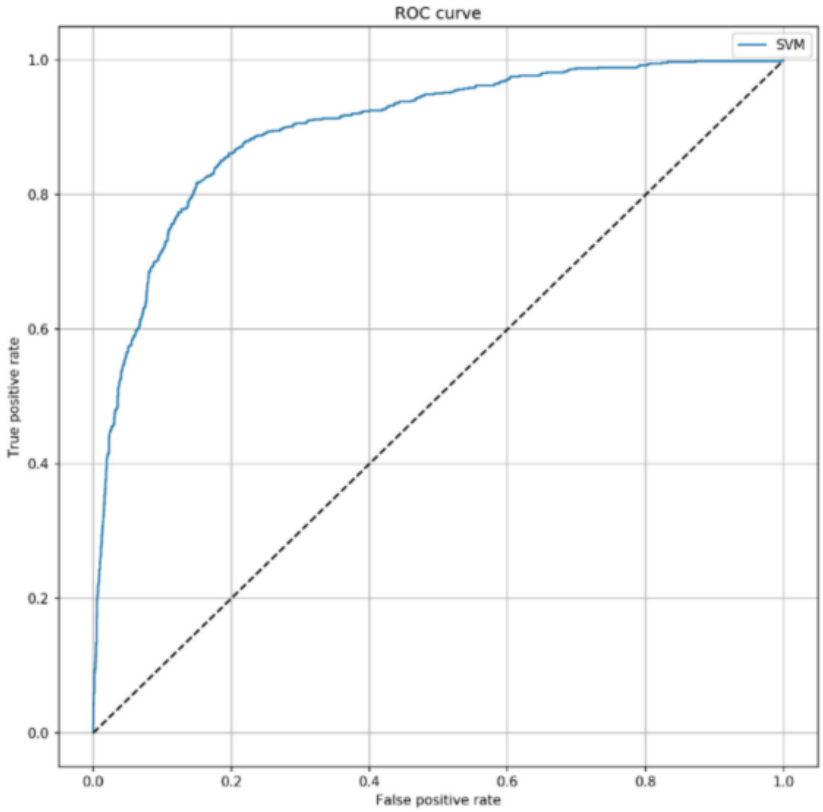

**The SVM algorithm improved the prediction of patients who died** Next, we aimed to develop a machine learning algorithm, which could improve the prediction of 7-day mortality in ICU. To do that, we used the SVM algorithm by combining SAPS II with other characteristics collected at ICU admission. Interestingly, the ROC curve of SVM predictive model (figure 6.12) achieved an AUC of 0.896 (95%CI = $[0.881, 0.910]$; $p < 0.001$), with an accuracy of 83.5% (95%CI = $[82.4\%, 84.7\%]$). In line, sensitivity and specificity were 81.0% (95%CI = $[79.9\%, 82.1\%]$) and 84.0% (95%CI = $[82.9\%, 85.1\%]$), respectively.

Figure 6.12: ROC curve of the SVM algorithm to predict 7-day mortality




**The SVM algorithm maintained its predictive ability among patients who did not develop HAIs** We also tested the predictive ability of the SVM classifier among patients who did not develop HAIs within 7 days from ICU admission. To do that, we removed 520 patients with at least one HAIs from the test set. Interestingly, the model did not depend on the onset of HAI, since both AUC (0.903; 95%CI = $[0.881, 0.912]$; $p < 0.001$) and accuracy (83.8%; 95%CI = $[82.6\%, 85.0\%]$) remained stable (Figure 3). In line, sensitivity and specificity were comparable to those obtained in the overall analysis (82.0%; 95%CI = $[80.8\%, 83.2\%]$; and 84.0%; 95%CI = $[82.8\%, 85.2\%]$, respectively).

Figure 6.13: ROC curve of the SVM algorithm to predict 7-day mortality, by excluding infected patients




**The predictivie performance of the SVM model by removing SAPS II**

The Shapley plot reported in figure 6.14 shows the contribution of each predictors to the SVM model output in terms of SHAP value. SHAP, which stands for *Shapley Additive exPlanations*, is an interpretability method based on Shapley values, a solution concept in cooperative game theory named in honor of Lloyd Shapley, who introduced it in 1951 [104] and won the Nobel Prize in Economics for it in 2012. SHAP was introduced by Lundberg and Lee (2017) [105] to explain individual predictions of any machine learning model. The explanation model is represented by a linear model — an additive feature attribution method — or just the summation of present features in the coalition game.

Since SAPS II was the predictor with the highest importance, we finally evaluated the predictive performance of the classifier after removing SAPS II. Interestingly, the SVM model without SAPS II led to an AUC of 0.653 (95%CI = $[0.632, 0.675]$;

$p < 0.001$), with an accuracy of 68.4% (95%CI = [66.9%, 69.8%]) on the test set (figure 6.15). Accordingly, sensitivity and specificity decreased to 32.0% (95%CI = [30.5%, 33.5%]) and 74.0% (95%CI = [72.5%, 75.5%]), respectively.

Figure 6.14: Shapley plot showing the contribution of each predictor to the SVM model output

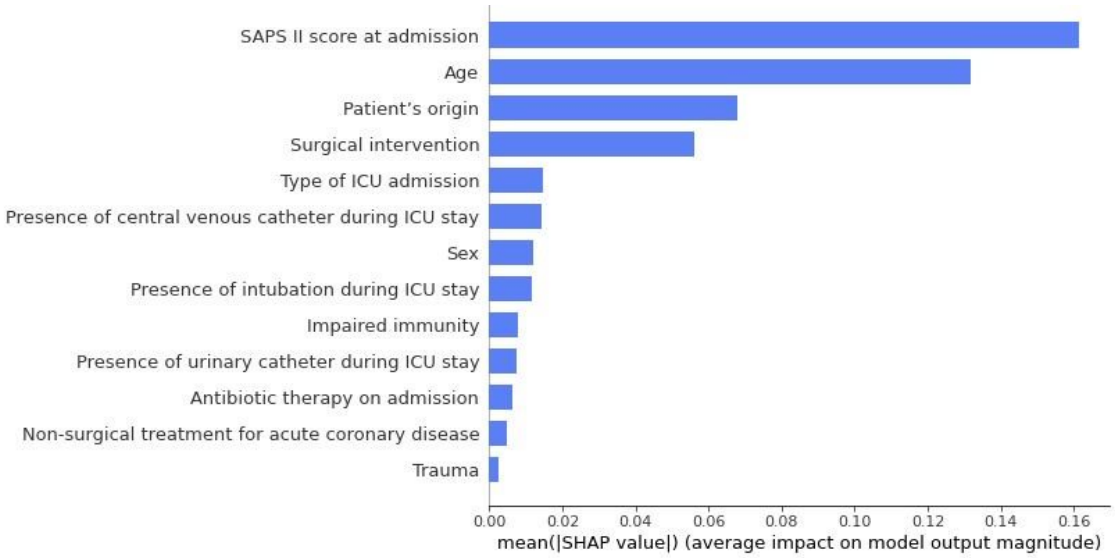
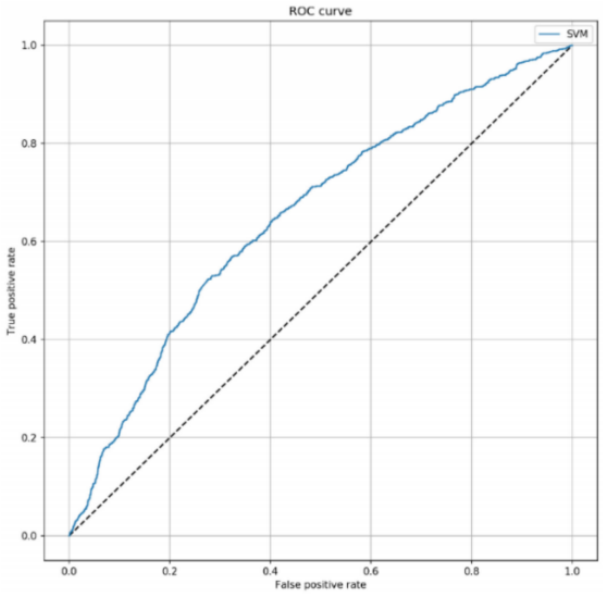



Figure 6.15: ROC curve of SVM algorithm predicting 7-day mortality, by excluding SAPS II score



### Discussion

In past years, numerous early warning scores were developed and employed to monitor and predict patients conditions and severity, as well as their adverse events in healthcare facilities [6–9]. Among these, SAPS II still represents one of the most widely used tool to estimate patients risk of death and other adverse outcomes [2, 10–16, 84, 106]. In view of these considerations, we first aimed to evaluate the accuracy of SAPS II alone to identify patients who died within seven days from their admission in ICUs, using a large dataset from the SPIN-UTI project. Although AUC obtained was statistically significant, however, the low accuracy of nearly 69% discouraged the routinely application of SAPS II to achieve this purpose.

Applying novel predictive algorithms, however, could be important to ameliorate patients safety and management in clinical practice, especially in the ICU setting. Thus, we hypothesized that combining SAPS II with other variables collected at ICU admission could improve the prediction of 7-day mortality [97, 98, 103, 107, 108]. Indeed, its now well-established that machine learning algorithms could overcome the limitations of traditional existing tools, also allowing early prediction of mortality [6–9, 18–21, 107, 109]. To do that, we developed a SVM model, which combined SAPS II with the following patients characteristics at ICU admission: sex, patients origin, type of ICU admission, non-surgical treatment for acute coronary disease, surgical intervention, presence of invasive devices at ICU admission, trauma, impaired immunity, antibiotic therapy in 48 hours before or after ICU admission and presence of infection in seven days of ICU stay. The model exhibited an AUC of 0.90 with an accuracy of 83.5% on the test set. Interestingly, its predictive performance was higher than SAPS II alone and even than a logistic regression model including additional patients characteristics collected at ICU admission. We also demonstrated that the performance in predicting 7-day mortality was also similar in only patients who did not acquired HAIs during their hospitalization.

Overall, our findings underlined the potentially crucial role of machine learning algorithms in many public health issues, providing clinicians with better diagnostic tools and improving medical care in the next future. The promising benefits of applying machine learning on healthcare quality rely on the opportunity of making prevention and diagnosis as early as possible, in a context of precision medicine applicable to all settings. Indeed, these algorithms if properly applied could overcome

limitations of existing traditional early warning scores 43-46, 54-60. For instance, a previous study developed a machine learning algorithm based on vital signs of ICU patients, such as heart and respiration rate, oxygen saturation and blood pressure. In particular, the algorithm was able to predict mortality in ICU with an accuracy of 91.6% REF. Our findings - together with those from other research groups lay the foundations to develop automated and real-time tools able to identify patients who need more attention because of their high risk of death.

Our model had several strengths, including the better ability of predicting 7-day mortality if compared with an early warning score as SAPS II. Moreover, our model was trained and tested on large datasets obtained through patient-based surveillance, structured and standardized according to ECDC protocol. This should allow validation and comparison with other European countries. On the other hand, however, we cannot completely exclude historical bias due to a 14-year period of data collection. Beyond that, there were other considerations to keep in mind when interpreting our results. The first one was that our findings confirmed the importance of developing and validating early warning scores to predict the risk of death and other adverse outcomes in ICUs and other wards. Indeed, although we used several variables collected at ICU admission, the removal of SAPS II from the model significantly reduced the predictive performance. The second consideration was that machine learning requires a lot of variables and records, which are not always available in each healthcare settings. Although we used variables that can be easily collectible at ICU admission (e.g., patients demographic, origin and type of admission, medical history, and disease severity), a lot of patients had both structural missing and missing at random data. While the first type of missing data could be easily managed by improving their collection in the next SPIN-UTI editions (e.g., making them mandatory), those that miss at random will continue to exist. This still remains a common issue encountered when analyzing real-world data. For this reason, we cannot completely exclude potential bias related to the high proportion of missing data. To partially manage missing data, we adopted a dual approach to generate synthetic records from those incomplete. Indeed, we created a dataset of synthetic records that was used as the training set for our machine learning algorithm. However, while it remains preferable using real data to train the algorithms,

the comparison between training and test sets showed no significant differences. Finally, we recognize that our model did not take into account the time component. Indeed, we used non-temporal variables related to patients characteristics at ICU admission and 7-day mortality as the main outcome. Thus, it will be our task to consider a time-series approach (e.g., survival analysis) for improving our model.

With these considerations in mind, to the best of our knowledge, our study is the first employing the SVM algorithm to discriminate patients who died within seven days from their ICU admission from those who did not. The model showed good predictive performance, even though improvable. For this reason, further studies should be encouraged to develop and validate risk prediction models, which could help to predict adverse outcome as early as possible, and to improve patient care globally.

# Chapter 7

# Conclusions

It is known that machine learning and visual analytics techniques have been widely used in the healthcare area, including disease and risk prediction. It is also known that decision-making responses suffer, in the first place, from the class imbalance issue and, in the second place, from the general quality of the data in terms of missingness, bias and significance. These problems have received much attention from researchers in the fields of medical, fraud detection [110] and bankruptcy prediction [111]. To deal with these problems, advanced data-driven and machine learning techniques have been developed constantly.

The data collected in the field of public health over the years, as in the SPIN-UTI project, suffer particularly from the problems of completeness and balancing. This work proposes a series of methods to deal the problems of unbalanced and incomplete datasets together, as well as a tool for visualizing the data in question, as the result of longitudinal observational studies. The developed visual analytics techniques prove that the quantitative visualization of data in the form of sequences of related events leads to a better understanding of the results. In addition, the described data augmentation methods are effective to improve machine learning calssifiers. In particular, the 1-NN algorithm for a multivariate data imputation, applied to subsets divided by outcome of interest, with different metrics and considering the amount of missing values of each feature of the dataset, results very useful to recover otherwise unusable data. Furthermore, this method used in conjunction with one of the best known synthetic oversampling algorithms, represents an application strategy that allow to generate data in line with real ones, as demonstrated by the case studies, with the presence of a few duplicate records, which were eliminated without significantly affecting the total number of reconstructed observations, avoiding bias and

overfitting issues.

The results of this work show the applicability of the proposed solutions useful for further improving the public health risk prediction with particular attention to the risk assessment of HAIs (which as reported by the WHO, the global burden of HAIs raises up to 15% among all hospitalized patients, with a proportion that achieves more than 30% in those who stay in ICUs [112–114]). These solutions can enhance research, surveillance, and intervention levels in public health, thus allowing to implement more effective evidence-based policies. Furthermore, the described methods can be used outside of public health and health surveillance domains, since the problems addressed in the course of the research are common to other fields and thus provide new insights for future works.

# Bibliography

[1] G. Favara, P. M. Riela, A. Maugeri, M. Barchitta, G. Gallo, and A. Agodi. "Risk of Pneumonia and Associated Outcomes in Intensive Care Unit: An Integrated Approach of Visual and Cluster Analysis". In: *2019 IEEE World Congress on Services (SERVICES)*. Vol. 2642-939X. 2019, pp. 289–294. DOI: 10.1109/SERVICES.2019.00083.

[2] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman. "APACHE II: a severity of disease classification system". In: *Crit Care Med* 13.10 (Oct. 1985), pp. 818–829.

[3] M. Garrouste-Orgeas, J. F. Timsit, M. Tafflet, B. Misset, J.-R. Zahar, L. Soufir, T. Lazard, S. Jamali, B. Mourvillier, Y. Cohen, A. D. Lassence, E. Azoulay, C. Cheval, A. Descorps-Declere, C. Adrie, M.-A. C. de Beauregard, and J. C. and. "Excess Risk of Death from Intensive Care Unit–Acquired Nosocomial Bloodstream Infections: A Reappraisal". In: *Clinical Infectious Diseases* 42.8 (Apr. 2006), pp. 1118–1126. DOI: 10.1086/500318. URL: https://doi.org/10.1086/500318.

[4] J.-L. Vincent. "International Study of the Prevalence and Outcomes of Infection in Intensive Care Units". In: *JAMA* 302.21 (Dec. 2009), p. 2323. DOI: 10.1001/jama.2009.1754. URL: https://doi.org/10.1001/jama.2009.1754.

[5] W. Wang, S. Zhu, Q. He, R. Zhang, Y. Kang, M. Wang, K. Zou, Z. Zong, and X. Sun. "Developing a Registry of Healthcare-Associated Infections at Intensive Care Units in West China: Study Rationale and Patient Characteristics". In: *Clinical Epidemiology* Volume 11 (Dec. 2019), pp. 1035–1045. DOI: 10.2147/clep.s226935. URL: https://doi.org/10.2147/clep.s226935.

[6] S. Gerry, T. Bonnici, J. Birks, S. Kirtley, P. S. Virdee, P. J. Watkinson, and G. S. Collins. "Early warning scores for detecting deterioration in adult

hospital patients: systematic review and critical appraisal of methodology". In: *BMJ* (May 2020), p. m1501. DOI: 10.1136/bmj.m1501. URL: https://doi.org/10.1136/bmj.m1501.

[7] T. A. Brennan. "Incidence of adverse events and negligence in hospitalized patients: results of the Harvard Medical Practice Study I". In: *Quality and Safety in Health Care* 13.2 (Apr. 2004), pp. 145–151. DOI: 10.1136/qshc.2002.003822. URL: https://doi.org/10.1136/qshc.2002.003822.

[8] C. Vincent. "Adverse events in British hospitals: preliminary retrospective record review". In: *BMJ* 322.7285 (Mar. 2001), pp. 517–519. DOI: 10.1136/bmj.322.7285.517. URL: https://doi.org/10.1136/bmj.322.7285.517.

[9] K. Hillman, P. Bristow, T. Chey, K. Daffurn, T. Jacques, S. Norman, G. Bishop, and G. Simmons. "Duration of life-threatening antecedents prior to intensive care admission". In: *Intensive Care Medicine* 28.11 (Nov. 2002), pp. 1629–1634. DOI: 10.1007/s00134-002-1496-y. URL: https://doi.org/10.1007/s00134-002-1496-y.

[10] J. Allyn, C. Ferdynus, M. Bohrer, C. Dalban, D. Valance, and N. Allou. "Simplified Acute Physiology Score II as Predictor of Mortality in Intensive Care Units: A Decision Curve Analysis". In: *PLOS ONE* 11.10 (Oct. 2016). Ed. by C. Lazzeri, e0164828. DOI: 10.1371/journal.pone.0164828. URL: https://doi.org/10.1371/journal.pone.0164828.

[11] M. Gilani, M. Razavi, and A. Azad. "A comparison of Simplified Acute Physiology Score II, Acute Physiology and Chronic Health Evaluation II and Acute Physiology and Chronic Health Evaluation III scoring system in predicting mortality and length of stay at surgical intensive care unit". In: *Nigerian Medical Journal* 55.2 (2014), p. 144. DOI: 10.4103/0300-1652.129651. URL: https://doi.org/10.4103/0300-1652.129651.

[12] F. Sadaka, C. EthmaneAbouElMaali, M. A. Cytron, K. Fowler, V. M. Javaux, and J. O'Brien. "Predicting Mortality of Patients With Sepsis: A Comparison of APACHE II and APACHE III Scoring Systems". In: *Journal of Clinical Medicine Research* 9.11 (2017), pp. 907–910. DOI: 10.14740/jocmr3083w. URL: https://doi.org/10.14740/jocmr3083w.

[13] A. Agodi, M. Barchitta, and F. Auxilia. "Epidemiology of intensive care unit-acquired sepsis in Italy: results of the SPIN-UTI network". In: *annali di igiene medicina preventiva e di comunità* 5 (Oct. 2018), pp. 15–21. ISSN: 1120-9135. DOI: 10.7416/ai.2018.2247. URL: https://doi.org/10.7416/ai.2018.2247.

[14] D. H. Beck, G. B. Smith, J. V. Pappachan, and B. Millar. "External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study". In: *Intensive Care Medicine* 29.2 (Jan. 2003), pp. 249–256. DOI: 10.1007/s00134-002-1607-9. URL: https://doi.org/10.1007/s00134-002-1607-9.

[15] J. R. L. Gall. "A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study". In: *JAMA: The Journal of the American Medical Association* 270.24 (Dec. 1993), pp. 2957–2963. DOI: 10.1001/jama.270.24.2957. URL: https://doi.org/10.1001/jama.270.24.2957.

[16] A. B. Nielsen, H.-C. Thorsen-Meyer, K. Belling, A. P. Nielsen, C. E. Thomas, P. J. Chmura, M. Lademann, P. L. Moseley, M. Heimann, L. Dybdahl, L. Spangsege, P. Hulsen, A. Perner, and S. Brunak. "Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records". In: *The Lancet Digital Health* 1.2 (2019), e78–e89. ISSN: 2589-7500. DOI: https://doi.org/10.1016/S2589-7500(19)30024-X. URL: https://www.sciencedirect.com/science/article/pii/S258975001930024X.

[17] G. Kong, K. Lin, and Y. Hu. "Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU". In: *BMC Medical Informatics and Decision Making* 20.1 (Oct. 2020). DOI: 10.1186/s12911-020-01271-2. URL: https://doi.org/10.1186/s12911-020-01271-2.

[18] A. Scardoni, F. Balzarini, C. Signorelli, F. Cabitza, and A. Odone. "Artificial intelligence-based tools to control healthcare associated infections: A systematic review of the literature". In: *Journal of Infection and Public Health* 13.8

(Aug. 2020), pp. 1061–1077. DOI: 10.1016/j.jiph.2020.06.006. URL: https://doi.org/10.1016/j.jiph.2020.06.006.

[19] J. P. Parreco, A. E. Hidalgo, A. D. Badilla, O. Ilyas, and R. Rattan. "Predicting central line-associated bloodstream infections and mortality using supervised machine learning". In: *Journal of Critical Care* 45 (June 2018), pp. 156–162. DOI: 10.1016/j.jcrc.2018.02.010. URL: https://doi.org/10.1016/j.jcrc.2018.02.010.

[20] R. C. Deo. "Machine Learning in Medicine". In: *Circulation* 132.20 (Nov. 2015), pp. 1920–1930. DOI: 10.1161/circulationaha.115.001593. URL: https://doi.org/10.1161/circulationaha.115.001593.

[21] R. C. Deo. "Machine Learning in Medicine". In: *Circulation* 142.16 (Oct. 2020), pp. 1521–1523. DOI: 10.1161/circulationaha.120.050583. URL: https://doi.org/10.1161/circulationaha.120.050583.

[22] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury. "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes". In: *BMC Medical Informatics and Decision Making* 10.1 (Mar. 2010). DOI: 10.1186/1472-6947-10-16. URL: https://doi.org/10.1186/1472-6947-10-16.

[23] J. M. Snowden, S. Rose, and K. M. Mortimer. "Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique". In: 173.7 (Mar. 2011), pp. 731–738. DOI: 10.1093/aje/kwq472. URL: https://doi.org/10.1093/aje/kwq472.

[24] C. Snijders, U. Matzat, and U.-D. Reips. ""Big Data" : Big Gaps of Knowledge in the Field of Internet Science". In: *International Journal of Internet Science* 7 (Jan. 2012), pp. 1–5.

[25] M. Cox and D. Ellsworth. "Application-controlled demand paging for out-of-core visualization". In: (Nov. 1997), pp. 235–244. DOI: 10.1109/VISUAL.1997.663888.

[26] D. Laney. *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Tech. rep. META Group, Feb. 2001. URL: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[27] C. Onay and E. Öztürk. "A review of credit scoring research in the age of Big Data". In: *Journal of Financial Regulation and Compliance* 26 (2018), pp. 382–405.

[28] C. Fox. *Data Science for Transport A Self-Study Guide with Computer Exercises*. eng. 1st ed. 2018. 2018. ISBN: 9783319729534. URL: http://lib.ugent.be/catalog/ebk01:4100000002485403.

[29] M. Barchitta, A. Maugeri, G. Favara, P. M. Riela, G. Gallo, I. Mura, and A. Agodi. "Early Prediction of Seven-Day Mortality in Intensive Care Unit Using a Machine Learning Model: Results from the SPIN-UTI Project". In: *Journal of Clinical Medicine* 10.5 (2021). ISSN: 2077-0383. DOI: 10.3390/jcm10050992. URL: https://www.mdpi.com/2077-0383/10/5/992.

[30] T. Murdoch and A. Detsky. "The Inevitable Application of Big Data to Health Care". In: *JAMA : the journal of the American Medical Association* 309 (Apr. 2013), pp. 1351–2. DOI: 10.1001/jama.2013.393.

[31] M. Barchitta, A. Maugeri, G. Favara, P. M. Riela, C. La Mastra, M. La Rosa, R. M. San Lio, G. Gallo, I. Mura, A. Agodi, and SPIN-UTI Network. "Cluster analysis identifies patients at risk of catheter-associated urinary tract infections in intensive care units: findings from the SPIN-UTI Network". In: *Journal of Hospital Infection* 107 (2021), pp. 57–63. ISSN: 0195-6701. DOI: https://doi.org/10.1016/j.jhin.2020.09.030. URL: https://www.sciencedirect.com/science/article/pii/S0195670120304552.

[32] T. E. Raghunathan. "What do we do with missing data? Some options for analysis of incomplete data". In: *Annu Rev Public Health* 25 (2004), pp. 99–117.

[33] J. Pearl. "Causal Inference in Statistics: An Overview". In: *Statistics Surveys* 3 (Jan. 2009), pp. 96–146. DOI: 10.1214/09-SS057.

[34] S. Morgan and C. Winship. *Counterfactuals and Causal inference*. eng. 1st ed. 2018. 2007. ISBN: 978-0-521-67193-4.

[35] J. Robins. "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". In: *Mathematical Modelling* 7.9 (1986), pp. 1393–1512. ISSN:

0270-0255. DOI: https://doi.org/10.1016/0270-0255(86)90088-6. URL: https://www.sciencedirect.com/science/article/pii/0270025586900886.

[36] D. B. Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of Educational Psychology* 66.5 (1974), pp. 688–701. DOI: 10.1037/h0037350. URL: https://doi.org/10.1037/h0037350.

[37] M. A. Hernan. "A definition of causal effect for epidemiological research". In: *Journal of Epidemiology & Community Health* 58.4 (Apr. 2004), pp. 265–271. DOI: 10.1136/jech.2002.006361. URL: https://doi.org/10.1136/jech.2002.006361.

[38] J. M. Snowden, S. Rose, and K. M. Mortimer. "Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique". In: *American Journal of Epidemiology* 173.7 (Mar. 2011), pp. 731–738. DOI: 10.1093/aje/kwq472. URL: https://doi.org/10.1093/aje/kwq472.

[39] W. S. "Correlation and causation". In: *Journal of Agricultural Research,* 20 (1921), pp. 557–585.

[40] D. OD. *Structural Equations Models*. New York, NY: Academic Press, 1974.

[41] J. Robins. "A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods". In: *Journal of Chronic Diseases* 40 (Jan. 1987), 139S–161S. DOI: 10.1016/s0021-9681(87)80018-8. URL: https://doi.org/10.1016/s0021-9681(87)80018-8.

[42] J. M. Robins. "Causal Inference from Complex Longitudinal Data". In: *Latent Variable Modeling and Applications to Causality*. Ed. by M. Berkane. New York, NY: Springer New York, 1997, pp. 69–117.

[43] S. Greenland, J. Pearl, and J. M. Robins. "Causal diagrams for epidemiologic research". In: *Epidemiology* 10.1 (Jan. 1999), pp. 37–48.

[44] K. Wongsuphasawat and D. Gotz. "Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (Dec. 2012), pp. 2659–2668. DOI: 10.1109/tvcg.2012.225. URL: https://doi.org/10.1109/tvcg.2012.225.

[45] M. H. P. R. Sankey. "Temperature Entropy Or [theta Phi] Chart for 1 Lb. of H2O: Giving Also Volumes, Absolute Pressures and Dryness Fractions, Internal Energy, Water Heat, and Total Heat in British Thermal Units". In: (Dec. 1898).

[46] A. B. W. Kennedy and H. R. Sankey. "the thermal efficiency of steam engines. report of the committee appointed to the council upon the subject of the definition of astandard or standards of thermal efficiency for steam engines: with an introductory note (including appendixes and plate at back of volume)". In: *Minutes of the Proceedings of the Institution of Civil Engineers* 134.1898 (Jan. 1898), pp. 278–312. DOI: 10.1680/imotp.1898.19100. URL: https://doi.org/10.1680/imotp.1898.19100.

[47] M. Schmidt. "The Sankey Diagram in Energy and Material Flow Management". In: *Journal of Industrial Ecology* 12.1 (Feb. 2008), pp. 82–94. DOI: 10.1111/j.1530-9290.2008.00004.x. URL: https://doi.org/10.1111/j.1530-9290.2008.00004.x.

[48] C. Suetens, K. Latour, T. Kärki, E. Ricchizzi, P. Kinross, M. L. Moro, B. Jans, S. Hopkins, S. Hansen, O. Lyytikäinen, J. Reilly, A. Deptula, W. Zingg, D. Plachouras, and D. L. M. and. "Prevalence of healthcare-associated infections, estimated incidence and composite antimicrobial resistance index in acute care hospitals and long-term care facilities: results from two European point prevalence surveys, 2016 to 2017". In: *Eurosurveillance* 23.46 (Nov. 2018). DOI: 10.2807/1560-7917.es.2018.23.46.1800516. URL: https://doi.org/10.2807/1560-7917.es.2018.23.46.1800516.

[49] J. Steinmann, A. Knaust, A. Moussa, J. Joch, A. Ahrens, H. D. Walmrath, T. F. Eikmann, and C. E. Herr. "Implementation of a novel on-ward computer-assisted surveillance system for device-associated infections in an

intensive care unit". In: *Int J Hyg Environ Health* 211.1-2 (Mar. 2008), pp. 192–199.

[50] P. Zarb, B. Coignard, J. Griskeviciene, A. Muller, V. Vankerckhoven, K. Weist, M. M. Goossens, S. Vaerenberg, S. Hopkins, B. Catry, D. L. Monnet, H. Goossens, C. Suetens, C. N. C. P. for the ECD, and C. H. C. P. for the ECD. "The European Centre for Disease Prevention and Control (ECDC) pilot point prevalence survey of healthcare-associated infections and antimicrobial use". In: *Eurosurveillance* 17.46 (Nov. 2012). DOI: 10.2807/ese.17.46.20316-en. URL: https://doi.org/10.2807/ese.17.46.20316-en.

[51] M.-L. Lambert, G. Silversmit, A. Savey, M. Palomar, M. Hiesmayr, A. Agodi, B. V. Rompaye, K. Mertens, and S. Vansteelandt. "Preventable Proportion of Severe Infections Acquired in Intensive Care Units: Case-Mix Adjusted Estimations from Patient-Based Surveillance Data". In: *Infection Control & Hospital Epidemiology* 35.5 (May 2014), pp. 494–501. DOI: 10.1086/675824. URL: https://doi.org/10.1086/675824.

[52] U. M. Devlin, B. A. McNulty, A. P. Nugent, and M. J. Gibney. "The use of cluster analysis to derive dietary patterns: methodological considerations, reproducibility, validity and the effect of energy mis-reporting". In: *Proceedings of the Nutrition Society* 71.4 (Aug. 2012), pp. 599–609. DOI: 10.1017/s0029665112000729. URL: https://doi.org/10.1017/s0029665112000729.

[53] E. C. for Disease Prevention and Control. *Antimicrobial consumption: annual epidemiological report for 2017*. 2018.

[54] S. Saint, J. Wiese, J. K. Amory, M. L. Bernstein, U. D. Patel, J. K. Zemencuk, S. J. Bernstein, B. A. Lipsky, and T. P. Hofer. "Are physicians aware of which of their patients have indwelling urinary catheters?" In: *The American journal of medicine* 109.6 (2000), pp. 476–480.

[55] W. E. Stamm, T. M. Hooton, J. R. Johnson, C. Johnson, A. Stapleton, P. L. Roberts, S. L. Moseley, and S. D. Fihn. "Urinary tract infections: from pathogenesis to treatment". In: *The Journal of infectious diseases* 159.3 (1989), pp. 400–406.

[56] A. L. Flores-Mireles, J. N. Walker, M. Caparon, and S. J. Hultgren. "Urinary tract infections: epidemiology, mechanisms of infection and treatment options". In: *Nature reviews microbiology* 13.5 (2015), pp. 269–284.

[57] D. C. Burton, J. R. Edwards, A. Srinivasan, S. K. Fridkin, and C. V. Gould. "Trends in catheter-associated urinary tract infections in adult intensive care units—United States, 1990–2007". In: *Infection Control & Hospital Epidemiology* 32.8 (2011), pp. 748–756.

[58] H. J. Yoon, J. Y. Choi, Y. S. Park, C. O. Kim, J. M. Kim, D. E. Yong, K. W. Lee, and Y. G. Song. "Outbreaks of Serratia marcescens bacteriuria in a neurosurgical intensive care unit of a tertiary care teaching hospital: a clinical, epidemiologic, and laboratory perspective". In: *American journal of infection control* 33.10 (2005), pp. 595–601.

[59] J. M. Galiczewski. "Interventions for the prevention of catheter associated urinary tract infections in intensive care units: an integrative review". In: *Intensive and critical care Nursing* 32 (2016), pp. 1–11.

[60] L. E. Nicolle. "Catheter associated urinary tract infections". In: *Antimicrobial resistance and infection control* 3.1 (2014), pp. 1–8.

[61] D. R. Schaberg, R. A. Weinstein, and W. E. Stamm. "Epidemics of nosocomial urinary tract infection caused by multiply resistant gram-negative bacilli: Epidemiology and control". In: *The Journal of infectious diseases* 133.3 (1976), pp. 363–366.

[62] C. V. Gould, C. A. Umscheid, R. K. Agarwal, G. Kuntz, D. A. Pegues, H. I. C. P. A. Committee, et al. "Guideline for prevention of catheter-associated urinary tract infections 2009". In: *Infection Control & Hospital Epidemiology* 31.4 (2010), pp. 319–326.

[63] J. Meddings and S. Saint. *Disrupting the life cycle of the urinary catheter.* 2011.

[64] C. Chenoweth and S. Saint. "Preventing catheter-associated urinary tract infections in the intensive care unit". In: *Critical care clinics* 29.1 (2013), pp. 19–32.

[65] S. Saint, M. T. Greene, S. L. Krein, M. A. Rogers, D. Ratz, K. E. Fowler, B. S. Edson, S. R. Watson, B. Meyer-Lucas, M. Masuga, et al. "A program to prevent catheter-associated urinary tract infection in acute care". In: *New England Journal of Medicine* 374.22 (2016), pp. 2111–2119.

[66] S. Saint and C. E. Chenoweth. "Biofilms and catheter-associated urinary tract infections". In: *Infectious Disease Clinics* 17.2 (2003), pp. 411–432.

[67] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., Aug. 2002. DOI: 10.1002/9781119013563. URL: https://doi.org/10.1002/9781119013563.

[68] C. M. Salgado, C. Azevedo, H. Proença, and S. M. Vieira. "Missing Data". In: *Secondary Analysis of Electronic Health Records*. Cham: Springer International Publishing, 2016, pp. 143–162. ISBN: 978-3-319-43742-2. DOI: 10.1007/978-3-319-43742-2_13. URL: https://doi.org/10.1007/978-3-319-43742-2_13.

[69] D. Polit. *Nursing research : principles and methods*. Philadelphia: Lippincott, 1983. ISBN: 9780397544035.

[70] D. B. Rubin, ed. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc., June 1987. DOI: 10.1002/9780470316696. URL: https://doi.org/10.1002/9780470316696.

[71] R. Malarvizhi and A. Thanamani. "K-nearest neighbor in missing data imputation". In: *International Journal of Engineering Research and Development* 5.1 (Nov. 2012), pp. 05–07. ISSN: 2268-800X.

[72] N. Japkowicz. "The Class Imbalance Problem: Significance and Strategies". In: *In Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI*. 2000, pp. 111–117.

[73] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16 (June 2002), pp. 321–357. DOI: 10.1613/jair.953. URL: https://doi.org/10.1613/jair.953.

[74] S. Pai and G. D. Bader. "Patient Similarity Networks for Precision Medicine". In: *Journal of Molecular Biology* 430.18 (Sept. 2018), pp. 2924–2938. DOI: 10.1016/j.jmb.2018.05.037. URL: https://doi.org/10.1016/j.jmb.2018.05.037.

[75] E. Parimbelli, S. Marini, L. Sacchi, and R. Bellazzi. "Patient similarity for precision medicine: A systematic review". In: *Journal of Biomedical Informatics* 83 (July 2018), pp. 87–96. DOI: 10.1016/j.jbi.2018.06.001. URL: https://doi.org/10.1016/j.jbi.2018.06.001.

[76] A. Agodi, M. Barchitta, A. Quattrocchi, E. Spera, G. Gallo, F. Auxilia, S. Brusaferro, M. M. D'Errico, M. T. Montagna, C. Pasquarella, S. Tardivo, and I. M. and. "Preventable proportion of intubation-associated pneumonia: Role of adherence to a care bundle". In: *PLOS ONE* 12.9 (Sept. 2017). Ed. by Y. R. Kou, e0181170. DOI: 10.1371/journal.pone.0181170. URL: https://doi.org/10.1371/journal.pone.0181170.

[77] A. Agodi, F. Auxilia, M. Barchitta, S. Brusaferro, M. D'Errico, M. Montagna, C. Pasquarella, S. Tardivo, and I. Mura. "Antibiotic consumption and resistance: results of the SPIN-UTI project of the GISIO-SItI". In: *Epidemiologia e prevenzione* 39 (Oct. 2015), pp. 94–98.

[78] A. Agodi, F. Auxilia, M. Barchitta, S. Brusaferro, D. D'Alessandro, O. Grillo, M. Montagna, C. Pasquarella, E. Righi, S. Tardivo, V. Torregrossa, and I. Mura. "Trends, risk factors and outcomes of healthcare-associated infections within the Italian network SPIN-UTI". In: *Journal of Hospital Infection* 84.1 (May 2013), pp. 52–58. DOI: 10.1016/j.jhin.2013.02.012. URL: https://doi.org/10.1016/j.jhin.2013.02.012.

[79] A. Agodi, F. Auxilia, M. Barchitta, S. Brusaferro, D. D'Alessandro, M. Montagna, G. Orsi, C. Pasquarella, V. Torregrossa, C. Suetens, and I. Mura. "Building a benchmark through active surveillance of intensive care unit-acquired infections: the Italian network SPIN-UTI". In: *Journal of Hospital Infection* 74.3 (Mar. 2010), pp. 258–265. DOI: 10.1016/j.jhin.2009.08.015. URL: https://doi.org/10.1016/j.jhin.2009.08.015.

[80] A. Agodi, F. Auxilia, M. Barchitta, M. M. D'Errico, M. T. Montagna, C. Pasquarella, S. Tardivo, and I. Mura. "[Control of intubator associated pneumonia in intensive care unit: results of the GISIO-SItI SPIN-UTI Project]". In: *Epidemiol Prev* 38.6 Suppl 2 (2014), pp. 51–56.

[81] V. Agodi, M. Barchitta, and I. Mura. "The commitment of the GISIO-SItI to contrast Healthcare-Associated Infections and the experience of prevalence studies in Sicily". In: *annali di igiene medicina preventiva e di comunità* 4 (Aug. 2018), pp. 38–47. ISSN: 1120-9135. DOI: 10.7416/ai.2018.2233. URL: https://doi.org/10.7416/ai.2018.2233.

[82] "European surveillance of healthcare-associated infections in intensive care units- HAI-Net ICU protocol- Protocol version 1.02. ." In: (). URL: https://www.ecdc.europa.eu/sites/default/files/media/en/publications/Publications/healthcare-associated-infections-HAI-ICU-protocol.pdf.

[83] M. Barchitta, A. Maugeri, G. Favara, P. M. Riela, G. Gallo, I. Mura, and A. Agodi. "A machine learning approach to predict healthcare-associated infections at intensive care unit admission: findings from the SPIN-UTI project". In: *Journal of Hospital Infection* 112 (2021), pp. 77–86. ISSN: 0195-6701. DOI: https://doi.org/10.1016/j.jhin.2021.02.025. URL: https://www.sciencedirect.com/science/article/pii/S0195670121000840.

[84] İ. Mungan, Ş. Bektaş, M. Altınkaya Çavuş, S. Sarı, and S. Turan. "The predictive power of SAPS-3 and SOFA scores and their relations with patient outcomes in the Surgical Intensive Care Unit". In: *Turk J Surg* 35.2 (June 2019), pp. 124–130.

[85] M.-B. F. D, L.-M. Hilev, L.-P. David, R.-L. J. C, O.-R. Versis, and M.-A. J. L. "Performance of Three Prognostic Models in Critically Ill Patients with Cancer: A Prospective Study". In: *International Journal of Cancer and Clinical Research* 6.3 (July 2019). DOI: 10.23937/2378-3419/1410120. URL: https://doi.org/10.23937/2378-3419/1410120.

[86] C. Cortes and V. Vapnik. "Support-vector networks". In: *Machine Learning* 20.3 (Sept. 1995), pp. 273–297. DOI: 10.1007/bf00994018. URL: https://doi.org/10.1007/bf00994018.

[87] S.-j. Han, C. Qubo, and H. Meng. "Parameter selection in SVM with RBF kernel function". In: *World Automation Congress 2012* (2012), pp. 1–4.

[88] C. R. Yee, N. R. Narain, V. R. Akmaev, and V. Vemulapalli. "A Data-Driven Approach to Predicting Septic Shock in the Intensive Care Unit". In: *Biomedical Informatics Insights* 11 (Jan. 2019), p. 117822261988514. DOI: 10.1177/1178222619885147. URL: https://doi.org/10.1177/1178222619885147.

[89] L. Chen, A. Dubrawski, D. Wang, M. Fiterau, M. Guillame-Bert, E. Bose, A. M. Kaynar, D. J. Wallace, J. Guttendorf, G. Clermont, M. R. Pinsky, and M. Hravnak. "Using Supervised Machine Learning to Classify Real Alerts and Artifact in Online Multisignal Vital Sign Monitoring Data". In: *Critical Care Medicine* 44.7 (July 2016), e456–e463. DOI: 10.1097/ccm.0000000000001660. URL: https://doi.org/10.1097/ccm.0000000000001660.

[90] M. M. Churpek, T. C. Yuen, C. Winslow, D. O. Meltzer, M. W. Kattan, and D. P. Edelson. "Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards". In: *Critical Care Medicine* 44.2 (Feb. 2016), pp. 368–374. DOI: 10.1097/ccm.0000000000001571. URL: https://doi.org/10.1097/ccm.0000000000001571.

[91] N. Peiffer-Smadja, T. Rawson, R. Ahmad, A. Buchard, P. Georgiou, F.-X. Lescure, G. Birgand, and A. Holmes. "Corrigendum to 'machine learning for clinical decision support in infectious diseases: a narrative review of current applications' clinical microbiology and infection (2020) 584-595". In: *Clinical Microbiology and Infection* 26.8 (Aug. 2020), p. 1118. DOI: 10.1016/j.cmi.2020.05.020. URL: https://doi.org/10.1016/j.cmi.2020.05.020.

[92] A. Vellido, V. Ribas, C. Morales, A. R. Sanmartín, and J. C. R. Rodríguez. "Machine learning in critical care: state-of-the-art and a sepsis case study". In: *BioMedical Engineering OnLine* 17.S1 (Nov. 2018). DOI: 10.1186/s12938-018-0569-2. URL: https://doi.org/10.1186/s12938-018-0569-2.

[93] A. Ripoli, E. Sozio, F. Sbrana, G. Bertolino, C. Pallotto, G. Cardinali, S. Meini, F. Pieralli, A. M. Azzini, E. Concia, B. Viaggi, and C. Tascini. "Personalized machine learning approach to predict candidemia in medical wards".

In: *Infection* 48.5 (Aug. 2020), pp. 749–759. DOI: 10.1007/s15010-020-01488-3. URL: https://doi.org/10.1007/s15010-020-01488-3.

[94] B. Y. Li, J. Oh, V. B. Young, K. Rao, and J. Wiens. "Using Machine Learning and the Electronic Health Record to Predict Complicated Clostridium difficile Infection". In: *Open Forum Infectious Diseases* 6.5 (Apr. 2019). DOI: 10.1093/ofid/ofz186. URL: https://doi.org/10.1093/ofid/ofz186.

[95] N. Macesic, F. Polubriaginof, and N. P. Tatonetti. "Machine learning". In: *Current Opinion in Infectious Diseases* 30.6 (Dec. 2017), pp. 511–517. DOI: 10.1097/qco.0000000000000406. URL: https://doi.org/10.1097/qco.0000000000000406.

[96] J. A. Roth, M. Battegay, F. Juchler, J. E. Vogt, and A. F. Widmer. "Introduction to Machine Learning in Digital Healthcare Epidemiology". In: *Infection Control & Hospital Epidemiology* 39.12 (Nov. 2018), pp. 1457–1462. DOI: 10.1017/ice.2018.265. URL: https://doi.org/10.1017/ice.2018.265.

[97] M. Komorowski. "Artificial intelligence in intensive care: are we there yet?" In: *Intensive Care Medicine* 45.9 (June 2019), pp. 1298–1300. DOI: 10.1007/s00134-019-05662-6. URL: https://doi.org/10.1007/s00134-019-05662-6.

[98] A. Rajkomar, J. Dean, and I. Kohane. "Machine Learning in Medicine". In: *New England Journal of Medicine* 380.14 (Apr. 2019), pp. 1347–1358. DOI: 10.1056/nejmra1814259. URL: https://doi.org/10.1056/nejmra1814259.

[99] D. Linnen. "Statistical Modeling and Aggregate-Weighted Scoring Systems in Prediction of Mortality and ICU Transfer: A Systematic Review". In: *Journal of Hospital Medicine* 14.3 (2019), p. 161. DOI: 10.12788/jhm.3151. URL: https://doi.org/10.12788/jhm.3151.

[100] J.-P. Marcel, M. Alfa, F. Baquero, J. Etienne, H. Goossens, S. Harbarth, W. Hryniewicz, W. Jarvis, M. Kaku, R. Leclercq, S. Levy, D. Mazel, P. Nercelles, T. Perl, D. Pittet, C. Vandenbroucke-Grauls, N. Woodford, and V. Jarlier. "Healthcare-associated infections: think globally, act locally". In: *Clinical Microbiology and Infection* 14.10 (Oct. 2008), pp. 895–907. DOI: 10.1111/j.1469-0691.2008.02074.x. URL: https://doi.org/10.1111/j.1469-0691.2008.02074.x.

[101] D. Tan, T. Wiseman, V. Betihavas, and K. Rolls. "Patient, provider, and system factors that contribute to health care–associated infection and sepsis development in patients after a traumatic injury: An integrative review". In: *Australian Critical Care* 34.3 (May 2021), pp. 269–277. DOI: 10.1016/j.aucc.2020.08.004. URL: https://doi.org/10.1016/j.aucc.2020.08.004.

[102] E. Girou, M. Pinsard, I. Auriant, and M. Canonne. "Influence of the severity of illness measured by the Simplified Acute Physiology Score (SAPS) on occurrence of nosocomial infections in ICU patients". In: *Journal of Hospital Infection* 34.2 (Oct. 1996), pp. 131–137. DOI: 10.1016/s0195-6701(96)90138-3. URL: https://doi.org/10.1016/s0195-6701(96)90138-3.

[103] K. STRAND and H. FLAATTEN. "Severity scoring in the ICU: a review". In: *Acta Anaesthesiologica Scandinavica* 52.4 (Mar. 2008), pp. 467–478. DOI: 10.1111/j.1399-6576.2008.01586.x. URL: https://doi.org/10.1111/j.1399-6576.2008.01586.x.

[104] L. S. Shapley. *Notes on the N-Person Game - II: The Value of an N-Person Game*. Santa Monica, CA: RAND Corporation, 1951. DOI: 10.7249/RM0670.

[105] S. M. Lundberg and S.-I. Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[106] H. M., M. Lemdani, M. Gainier, H. Hubert, J. Tagne, and P. Lafaye de Micheaux. "Comparing the APACHE II, SOFA, LOD and SAPS II scores in patients who have developed a nosocomial infection". In: *Bangladesh Critical Care Journal* 2 (Mar. 2014), pp. 4–9. DOI: 10.3329/bccj.v2i1.19949.

[107] C. A. Lovejoy, V. Buch, and M. Maruthappu. "Artificial intelligence in the intensive care unit". In: *Critical Care* 23.1 (Jan. 2019). DOI: 10.1186/s13054-018-2301-9. URL: https://doi.org/10.1186/s13054-018-2301-9.

[108] C. Meiring, A. Dixit, S. Harris, N. S. MacCallum, D. A. Brealey, P. J. Watkinson, A. Jones, S. Ashworth, R. Beale, S. J. Brett, M. Singer, and A. Ercole. "Optimal intensive care outcome prediction over time using machine learning". In: *PLOS ONE* 13.11 (Nov. 2018). Ed. by L. A. Celi, e0206862. DOI: 10.1371/journal.pone.0206862. URL: https://doi.org/10.1371/journal.pone.0206862.

[109] T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh, D. Shimabukuro, U. Chettipally, M. D. Feldman, C. Barton, D. J. Wales, and R. Das. "Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach". In: *JMIR Medical Informatics* 4.3 (Sept. 2016), e28. DOI: 10.2196/medinform.5909. URL: https://doi.org/10.2196/medinform.5909.

[110] A. D. Pozzolo, O. Caelen, Y.-A. L. Borgne, S. Waterschoot, and G. Bontempi. "Learned lessons in credit card fraud detection from a practitioner perspective". In: *Expert Systems with Applications* 41.10 (Aug. 2014), pp. 4915–4928. DOI: 10.1016/j.eswa.2014.02.026. URL: https://doi.org/10.1016/j.eswa.2014.02.026.

[111] T. Le, M. Lee, J. Park, and S. Baik. "Oversampling Techniques for Bankruptcy Prediction: Novel Features from a Transaction Dataset". In: *Symmetry* 10.4 (Mar. 2018), p. 79. DOI: 10.3390/sym10040079. URL: https://doi.org/10.3390/sym10040079.

[112] P. Sulzgruber, S. Schnaubelt, L. Koller, G. Laufer, A. Pilz, N. Kazem, M.-P. Winter, B. Steinlechner, M. Andreas, T. Fleck, K. Distelmaier, G. Goliasch, A. Toma, C. Hengstenberg, and A. Niessner. "An Extended Duration of the Pre-Operative Hospitalization is Associated with an Increased Risk of Healthcare-Associated Infections after Cardiac Surgery". In: *Scientific Reports* 10.1 (May 2020). DOI: 10.1038/s41598-020-65019-8. URL: https://doi.org/10.1038/s41598-020-65019-8.

[113] E. Zimlichman, D. Henderson, O. Tamir, C. Franz, P. Song, C. K. Yamin, C. Keohane, C. R. Denham, and D. W. Bates. "Health Care–Associated Infections". In: *JAMA Internal Medicine* 173.22 (Dec. 2013), p. 2039. DOI:

10.1001/jamainternmed.2013.9763. URL: https://doi.org/10.1001/jamainternmed.2013.9763.

[114] "Report on the Burden of Endemic Health Care-Associated Infection Worldwide". In: (2011). URL: https://apps.who.int/iris/bitstream/handle/10665/80135/9789241501507_eng.pdf;sequence=1.