Prediction of mortality in septic patients with hypotension

Louis Mayaud

Saint Hilda's College

Department of Engineering Science

Supervised by
Prof. Lionel Tarassenko
Prof. Djillali Annane
Dr. Gari Clifford

Submitted: Hilary Term, 2014

This thesis is submitted to the Department of Engineering Science, University of Oxford, in fulfilment of the requirements for the degree of Doctor of Philosophy

Prediction of mortality in septic patients with hypotension

Louis Mayaud
Saint Hilda's College

Doctor of Philosophy
Hilary Term, 2014

ABSTRACT

Sepsis remains the second largest killer in the Intensive Care Unit (ICU), giving rise to a significant economic burden ($17b per annum in the US, 0.3% of the gross domestic product). The aim of the work described in this thesis is to improve the estimation of severity in this population, with a view to improving the allocation of resources.

A cohort of 2,143 adult patients with sepsis and hypotension was identified from the MIMIC-II database (v2.26). The implementation of state-of-the-art models confirms the superiority of the APACHE-IV model (AUC=73.3%) for mortality prediction using ICU admission data. Using the same subset of features, state-of-the art machine learning techniques (Support Vector Machines and Random Forests) give equivalent results. More recent mortality prediction models are also implemented and offer an improvement in discriminatory power (AUC=76.16%).

A shift from expert-driven selection of variables to objective feature selection techniques using all available covariates leads to a major gain in performance (AUC=80.4%). A framework allowing simultaneous feature selection and parameter pruning is developed, using a genetic algorithm, and this offers similar performance. The model derived from the first 24 hours in the ICU is then compared with a "dynamic" model derived over the same time period, and this leads to a significant improvement in performance (AUC=82.7%). The study is then repeated using data surrounding the hypotensive episode in an attempt to capture the physiological response to hypotension and the effects of treatment. A significant increase in performance (AUC=85.3%) is obtained with the static model incorporating data both before and after the hypotensive episode. The equivalent dynamic model does not demonstrate a statistically significant improvement (AUC=85.6%).

Testing on other ICU populations with sepsis is needed to validate the findings of this thesis, but the results presented in it highlight the role that data mining will increasingly play in clinical knowledge generation.

Acknowledgments

First of all, I would like to thank Prof. Djillali Annane for his unflinching support and the confidence he has placed in me over the past four years. I would also like to express my respect and gratitude to Prof. Lionel Tarassenko for the attentive supervision he honours his students with, and for having made no exception for me. Finally, a sincere thank you to Dr Gari Clifford for his enthusiasm for research, and the exemplary rigour of his scientific approach which is a good example to follow. I am deeply grateful to have worked with such significant people.

I would also like to acknowledge the tremendous scientific and moral support I received from my fellows students at the Institute of Biomedical Engineering (IBME), where I particularly appreciated the multi-cultural environment that I am going to miss terribly. Namely, I would like to thank Samuel Hugueny for the countless conversations we had late at night, and Alistair Johnson for being such a stimulating colleague and supportive friend, with whom I had the rare privilege of drinking good old ales in splendid places such as Boston, London, Paris, and Budapest.

On the other side of the Atlantic, I would like to thanks Prof. Roger Mark for hosting me on regular occasions in his lab. It has been a great honour to participate in the field of critical care data-mining and to interact with his group. Dr Leo Anthony Celi greatly contributed to my work thanks to his extensive knowledge of critical care practice and the vision he promotes. And of course, Dan, Ikaro, and Shamim with whom my scientific activities include transfer entropy calculation on the Charles river and switching Kalman filters applied to Latin dances.

Across the channel, I would like to thank all the people at Raymond Poincaré hospital with whom I have interacted over the past four years; in particular those who contributed to the RoBIK project. You allowed me to pursue this thesis. Firstly, I would like to thank Prof. David Orlikowski for supporting me financially. Also, Michèle, Marjorie, and Marine adopted me five years ago and have always been there when I needed them ever since. I would particularly like to thank Prof. Frederic Lofaso, Dr Hélène Prigent, and Dr Jérôme Aboab for their passion for physiology which they have shared with me. Working for the good of patients is something I learned in your company and is an example I mean to follow.

These four years of scientific adventure and labour between Oxford and Paris would simply not have been possible without the unswerving support of those who hosted me without condition: at any time of the day or the night, for an hour or three weeks, I always found the door open and your hospitality has always been impeccable. The time I spent in your company has been a great source of enjoyment and support, more then you would suspect. Estelle, Greg, Rodolphe, Anne-Gaëlle, Timothée, Solveig, Francois, Julia, and Louis-Marie: Merci! As soon as I have one, my house is yours.

To Dr Jacques Feldmar I would like to express my gratitude for being an illuminating example of a rare combination of excellence and humility, which I admire.

To the ones I love dearly, a simple and sincere thank you for teaching me everyday how to love “out of the box”. All the science of the world is worth nothing without it.

Contents

Introduction **1**

1 Background: sepsis, severe sepsis and septic shock **3**

1.1 History of sepsis 4

1.2 Key figures in the development of sepsis 6

1.2.1 The immune system 7

1.2.2 The cardiovascular system 8

1.2.3 The autonomous nervous system 11

1.3 Pathophysiology of sepsis 13

1.3.1 Step 1: Recognition of microbial-associated molecular patterns . . 14

1.3.2 Step 2: Pro and anti-inflammatory response 16

1.3.3 Step 3: Shock and organ failure 17

1.4 Clinical definitions of sepsis 18

1.4.1 Early recognition of infection and complication 18

1.4.2 The 1991 consensus conference 19

1.4.3 The 2001 definition 20

1.4.4 Latest definitions 21

1.4.5 Use of the definition of sepsis 21

1.5 Management of sepsis, severe sepsis and septic shock 22

1.5.1 Source identification and control 23

1.5.2 Haemodynamic support 24

1.5.3 Other supportive therapy . . . 30
1.6 Epidemiology and cost of sepsis . . . 31
1.6.1 Epidemiology of sepsis . . . 31
1.6.2 Costs of sepsis . . . 32
1.7 Estimation of sepsis severity . . . 33

2 Study population and data extracted 35
2.1 The MIMIC-II database . . . 35
2.1.1 An introduction the MIMIC-II database . . . 35
2.1.2 Dealing with de-identified medical data . . . 36
2.1.3 Database architecture . . . 37
2.1.4 Type of patients and services . . . 38
2.1.5 Description of available data . . . 39
2.1.6 Use of ontologies in the database . . . 40
2.2 Population Study . . . 41
2.2.1 Identification of the cohort of interest . . . 41
2.2.2 Description of the population studied . . . 45
2.3 Description of the data extracted . . . 46
2.3.1 Definition of outcome . . . 48
2.3.2 Physiological data . . . 49
2.3.3 Neurological status . . . 51
2.3.4 Microbiology results . . . 52
2.3.5 Medication and intervention . . . 52
2.3.6 Demographic data . . . 54
2.3.7 Chronic health conditions . . . 54
2.4 Pre-processing . . . 58

3 Baseline for prediction of mortality 60
3.1 Design, evaluation & comparison of severity scores . . . 60

3.1.1 Introducing simple modelling techniques . . . 61
3.1.2 Evaluation and comparison of model performance . . . 63
3.1.3 Comparison of model performance . . . 69
3.1.4 Choice of metric performance . . . 71
3.2 Existing strategies for the estimation of severity . . . 76
3.2.1 Previous attempts at predicting severity . . . 76
3.2.2 Other biomarkers of severity . . . 83
3.3 Performance on population of sepsis patients . . . 84
3.3.1 Performance reported in literature . . . 84
3.3.2 Performance on the sepsis population . . . 89
3.3.3 Discussion . . . 91
3.4 Conclusion . . . 93

4 Machine learning approach to prediction of mortality 95

4.1 Clinical data mining . . . 95
4.1.1 A short introduction to pattern recognition . . . 95
4.1.2 Machine learning modelling techniques . . . 98
4.1.3 Feature selection techniques . . . 105
4.2 State-of-the art modelling for the prediction of mortality: The Physionet challenge . . . 109
4.2.1 Bayesian Ensemble of forests . . . 110
4.2.2 Cascaded SVM-GLM paradigm . . . 112
4.3 Implementations on the severe sepsis population . . . 114
4.3.1 The data set . . . 114
4.3.2 Results . . . 116
4.4 Discussion . . . 117
4.5 Conclusion . . . 124

5 Predictive power of additional features derived from the data 126

5.1 Relevance or the estimation of variable importance . . . 127
5.1.1 Variable importance from random forest . . . 127
5.1.2 Variable importance for Bayesian ensemble of forests . . . 128
5.1.3 Other techniques for the estimation of variable importance . . . 128
5.2 Including additional co-variates . . . 129
5.3 Physiologically meaningful non-linear combination of raw variables . . . 130
5.3.1 Body indices . . . 130
5.3.2 Kidney indices . . . 131
5.3.3 Respiratory indices . . . 132
5.3.4 Cardiovascular indices . . . 133
5.3.5 Fluid balance and treatments . . . 134
5.3.6 Severity scores . . . 134
5.4 Values not missing at random . . . 135
5.5 Predictive power of the novel groups of variables . . . 135
5.5.1 Materials and methods . . . 135
5.5.2 Results . . . 137
5.5.3 Most predictive features . . . 139
5.6 Conclusion . . . 142

6 Feature selection and parameter pruning using a genetic algorithm 144
6.1 Introduction . . . 144
6.2 Description of Genetic Algorithm . . . 146
6.2.1 A brief introduction to Genetic Algorithms . . . 146
6.2.2 Initialization . . . 147
6.2.3 Iteration . . . 148
6.3 Different approaches to GA optimization . . . 152
6.3.1 Feature selection for logistic regression . . . 152
6.3.2 Parsimonious model inspired by automatic relevance determination 155

6.3.3 Feature selection and parameter pruning for support vector machines . . . 157
6.4 Results . . . 158
6.5 Discussion . . . 161

7 Adding dynamic information from clinical data to the prediction of mortality 167
7.1 Exploring the added value of dynamic information to improve the prediction of mortality . . . 169
7.1.1 Comparison technique . . . 169
7.1.2 Description of the datasets . . . 170
7.1.3 Results . . . 173
7.2 Relation between model accuracy and proximity to the outcome . . . 178
7.2.1 Comparison of models temporally closer to death . . . 180
7.2.2 Results . . . 181
7.3 Discussion . . . 182

8 Conclusion 189
8.1 Thesis overview . . . 189
8.2 Future work . . . 194
8.2.1 Feature extraction from waveforms . . . 194
8.2.2 Towards the digital ICU . . . 197
8.3 Conclusion . . . 200

Bibliography 201

Appendices 223

Acronyms 224

A Description of the variables 226

Introduction

Sepsis is the result of an inadequate response of the body to an infection. It is today the second largest killer in the Intensive Care Unit (ICU) after coronary disease [6] and its incidence is increasing slowly but steadily [189]. Simultaneously, the cost associated with the condition has been growing, giving both a clinical and an economic incentive to understand this disease better. Yet to date, the aetiology of sepsis remains only partly understood and its definition has seen major modifications over the past thirty years. During this period the number of scientific publications on this topic has been constantly increasing to about six thousand articles a year in 2012 (see figure 1), making this field one of the most important in medical research.

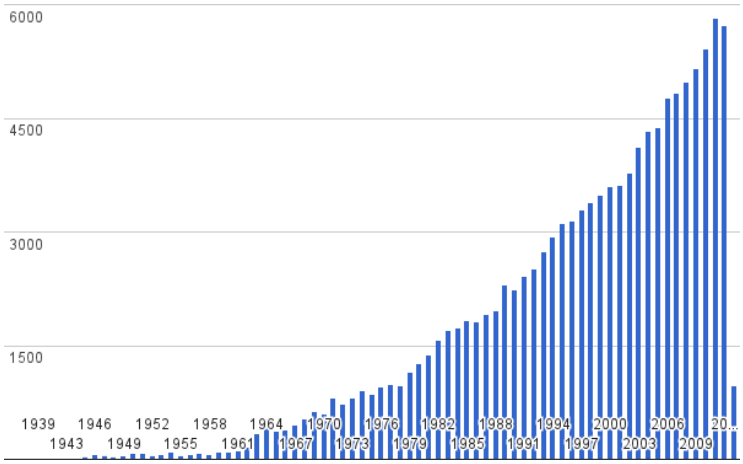


Figure 1: *Number of publications on sepsis since 1940: the number of hits in the PubMed database per year (x-axis) are shown by the blue bars.*

Predicting patients' mortality risk is an important field of biomedical research and remains a challenging task which has the objective of supporting clinicians in accurate risk stratification. Mortality can be assessed in many ways by considering ICU, in-hospital

or 28-day mortality for instance (see section 2.3.1). Current approaches focus on patient data (lab and haemodynamic values) at a point in time, to which simple modelling techniques are applied with respect to the defined outcome [16]. Although such techniques are relatively successful in describing global populations, they often fail to perform on homogeneous sub-populations and at the individual level [293]. For instance, patients experiencing sepsis-induced hypotension in the ICU share common physiological conditions: high temperature, elevated heart-rate (tachycardia), potentially signs of low blood pressure, altogether leading to a highly homogeneous population. Traditional mortality-prediction scores based on physiology, such as the Simplified Acute Physiology Score (SAPS), have been shown to have little predictive power for such a population [18, 22, 240]. This relatively poor performance can be explained by at least two factors:

- co-variates included in these severity scores have a low information content and other physiological variables should be considered for this subset of patients;
- the modelling techniques are too simple to represent the condition-specific patient data and state-of-the-art models should be evaluated instead.

In addition to this, severity scores generally do not account for caregiver intervention and consequently ignore a patient's response to them. Hence, these scores are invariably a static representation of the patient's physiology at admission (first 24 hours). Ferreira et al. [89] identified the amount of organ dysfunction as a strong predictor of mortality and suggested this variation may reflect the patient's response to treatment. Similarly, Bion et al. [29] demonstrated that the evolution of a physiological severity score was an independent predictor of mortality. Consequently, patients' dynamic information may provide additional information concerning their varying physiology, which could improve outcome prediction. More precisely, incorporating knowledge of treatments and of the subsequent physiological change may prove beneficial.

Guidelines for septic shock, where administration of treatments (fluids and pressors) is standardized and well localized in time, allow us to extract data from these events in order to explore our hypothesis.

Chapter 1

Background: sepsis, severe sepsis and septic shock

Introduction

Sepsis is a condition that develops when the physiological systems fail. The exact pathophysiology is not yet fully understood. However current knowledge has already successfully generated treatment recommendations [76] that have significantly reduced mortality in the last few decades [189, 6]. In order to provide a deeper understanding of the problem and to rationalise the scientific orientation of this work, we will first describe what is currently accepted on the aetiology of sepsis in section 1.3. To do so we will need to describe key components of human physiology (in sections 1.2.1 to 1.2.3), which will prove greatly beneficial to the understanding of both development and treatment of sepsis. Later, clinical considerations on the management of sepsis will be introduced (section 1.5) to anchor our work in bedside clinical reality. Finally, presenting the epidemiology of sepsis together with some economic data (section 1.7) will build a strong clinico-economical rationale for this research.

1.1 History of sepsis

The oldest known report of sepsis comes from a papyrus found in Luxor (Egypt) in the year 1862 and it was presumed to have been written around 1600 BC. The document is believed to be a copy of a manuscript from 3000 BC [42], which describes 48 cases of wounds. In five of them, fever is identified as a secondary phenomenon worsening prognosis. Similarly, the presence of pus was also recognized as a much later consequence of the initial insult and noted to be related to a worse outcome. As a consequence, Egyptian physicians limited surgical exploration so as to not promote lesion suppuration. Almost five thousand years before Semmelweis, Pasteur and Bone, the Egyptians [42] were aware of the concept of systemic response (fever) and inflammation (pus).

Unsurprisingly, Greek civilization has also contributed to the early recognition of sepsis. In fact, the word "sepsis" is first encountered in Homer's 24^{th} song of the Iliad [42] as $\sigma\eta\pi\omega$, which translates to "rotting" [103]:

> He is still just lying at the tents by the ship of Achilles, and though it is now twelve days that he has lain there, his flesh is not *wasted* nor have the worms eaten him although they feed on warriors.

This word is also present in Hippocrates's work (460-370 BC) and later literature: *sepsis* refers to putrefaction, rotten organic matter and is associated with a "bad" smell. The opposite concept of *pepsis* refers to fermentation, maturation and is associated with a positive outcome. As a consequence, the management of wounds and pus depends mostly on its attributes (colour and smell) and whether it can be categorised as *sepsis* or *pepsis*.

The Roman contribution to medical sciences was greatly influenced by Galen (129-200 AD) who vigorously promoted questioning of medical opinion and who, rather ironically, became the nearly unchallenged medical authority for the following fifteen centuries [226, 42]. His opinion on sepsis was unfortunately far from modern views as he claimed: "Pus bonum et laudabile", which led to fifteen hundred years of extensive use

of cautery and suppurative healing (promotion of pus formation by stimulation of the wound, for example by pouring hot oil onto the wound). A few individuals actually attempted to challenge this view; for example Henri de Mondeville (1260-1320) in Montpellier who suggested that wounds shall be cleaned with boiled water instead. Unfortunately such views remained mostly unnoticed.

Introduction of firearms in the second half of 14^{th} century exposed modern barber surgeons to non-haemorrhagic causes of death, raising new interest in the topic of infection. Paracelsus (1493-1541) virulently criticized Galen's work and claimed:

> The true physician of wounds is nature. All treatment must be reduced to infection prevention. Complexion, humors, diet and time, and the stars don't have any influence. The result will be determined by a treatment which allows nature to act in peace.

This five-hundred-year-old statement still resonates in modern critical care journals. The controversial use of Swan-Ganz catheters for instance may have been accountable for up to 25,000 deaths per year in the US [262]. In a less dramatic manner, excessive nocturnal interventions in the ward deteriorates patients' sleep with yet under-explored consequences on their recovery [210]. Similarly, it is unclear how harmful repeated blood sampling can be in a population of patients with hypovolaemia (low blood volume). Finally, the exact benefit of most ICU interventions for a specific patient relies at best on weak evidence [48]. Despite the large body of evidence suggesting that "less is more in the ICU" [185], resistance to change is still omnipresent. This resistance is similar to that which Pronovost et al. [228] faced when they suggested that the use of simple check-lists could reduce dramatically the incidence of infections related to cannulation. Unsurprisingly then, a recent article claimed that "half of patients do not receive optimal care" [157]; the large number of negative comments left by clinicians to the editor on the journal website however indicates that the medical community has evolved [261] towards more transparent and rational thinking. Unfortunately, resistance to change remains in critical care.

The empirical approach was reinstated almost by accident in the medical debate about sepsis with the misfortunes of Ambroise Paré (1509-1590). During the siege of Turin (1536), he was forced to interrupt traditional therapy, as oil ran out. Instead, he applied soothing ointment and rapidly noticed the improvement of his patients. The history of sepsis then meets one of the most tragic characters of modern science: Ignaz Semmelweis (1812-1865). He worked at one of Vienna's two obstetric clinics where the mortality of mothers from puerperal fever was Europe's highest. He identified physicians' hands as the vector of contamination and recommended hand washing between observations on cadavers and handling of women. He explored his hypothesis in a dramatically successful controlled study, generating a drop of mortality from 25% to below 5%. Despite the empirical evidence provided, political games in the higher spheres of the medical community completely discredited his work. An ironical twist of fate lead Semmelweis to die of sepsis in a mental institution where he spent his final months. It is interesting to know that his story was the topic of the medical dissertation by Louis Ferdinand Céline [53] - the famous and mostly controversial French author. The work of Louis Pasteur (1822-1895) later confirmed his findings and re-established his work with the discovery of micro-organisms and the invention of sterilization, both of which opened the way to modern medicine.

1.2 Key figures in the development of sepsis

Before describing the pathophysiology of sepsis, some of the physiological system involved during the onset and development of sepsis will be presented. In fact, the complexity of sepsis certainly is partly due to the interaction of the immune, cardio-vascular and autonomous nervous systems. This short introduction does not pretend to be exhaustive in its description of human physiology. However, the number and diversity of agents described in this section should provide a good overview of the mechanisms involved and aid the understanding of the following sections.

1.2.1 The immune system

The role of the immune system is to protect the organism against external microbial threats. It consists of a broad variety of biological structures and processes. The two most important functions of the immune system are: recognition of danger signals (such as *Non-Self*) and the destruction of structures exhibiting the associated molecular patterns. In addition to natural barriers (such as the skin) to external threats, the two major entities that protect the body are the *innate* and the *adaptive* immune systems.

The innate immune system handles the *non-specific* response of the body to an external threat and is largely mediated through the process of inflammation [151]. Following a harmful stimulus, or the recognition of external elements in the body, specialized cells are delivered to the site of injury in two steps: *local* vasodilation and increased permeability of blood vessels. The transfer of plasma from the blood to the tissue will result in swelling and redness that are typical of the inflammatory state. Because it plays a fundamental role in the early stages of sepsis, a more detailed description of this mechanism is provided in section 1.3.2.

The specialized cells involved in the immune response are called leukocytes or White Blood Cells (WBCs) when found in the blood stream. Their nomenclature is vast and the description here is restricted to the type playing a major role in the evolution of sepsis. Leukocytes are usually split into two groups depending on the role they play in the immune system:

1. Phagocytes are involved in the non-specific response of the innate system and are composed of different types of cells:

   (a) Granulocytes are small in size and are found in large quantities. They first identify bacteria or fungi they randomly meet in the body, attack them and alert more efficient cells such as macrophages. The granulocytes circulating in the blood are called neutrophils;

   (b) Macrophages are bigger cells designed to kill. Once recruited by granulo-

cytes, they will reach the site of infection, grow in number and kill. The broken molecular components of cells destroyed during phagocytosis (such as antigens) are internally processed by macrophages and presented to lymphocytes, the specialized cells of the immune system. Dendritic cells and Mast cells are other macrophages involved in the mediation of the adaptive response.

2. Lymphocytes are cells of the immune system coordinating the adaptive and specific response to an infection:

   (a) Thymus cells (T-cells) are activated by phagocytes and possess specific antibodies on their surface. These are recruited to suppress infection locally and simultaneously activate Bursa of Fabricius cells (B-cells);

   (b) B-cells subsequently produce large amounts of free circulating antibodies (in contrast to surface bound antibodies in T-cells) that will link to bacteria, disable them, and ease their recognition by macrophages.

The activation of the immune system as described in this section is mostly coordinated by small signalling proteins called *cytokines*. They are assumed to play a central role in the evolution of sepsis and will be covered in more detail in section 1.3.

1.2.2 The cardiovascular system

The primary function of the cardiovascular system is to deliver nutrients and oxygen in order to meet the tissues' metabolic demand. Simultaneously, by-products of cellular respiration (such as CO_2) are expelled to avoid toxicity. The cardiovascular system also supports other vital functions such as the immune and endocrine system by carrying various types of molecular agents (hormones and white cells for instance).

The circulation of blood from supply (lungs and intestines) to tissues is achieved by the action of the heart pumping blood through the vessels. As described in Figure 1.1, blood vessels are divided into two categories:

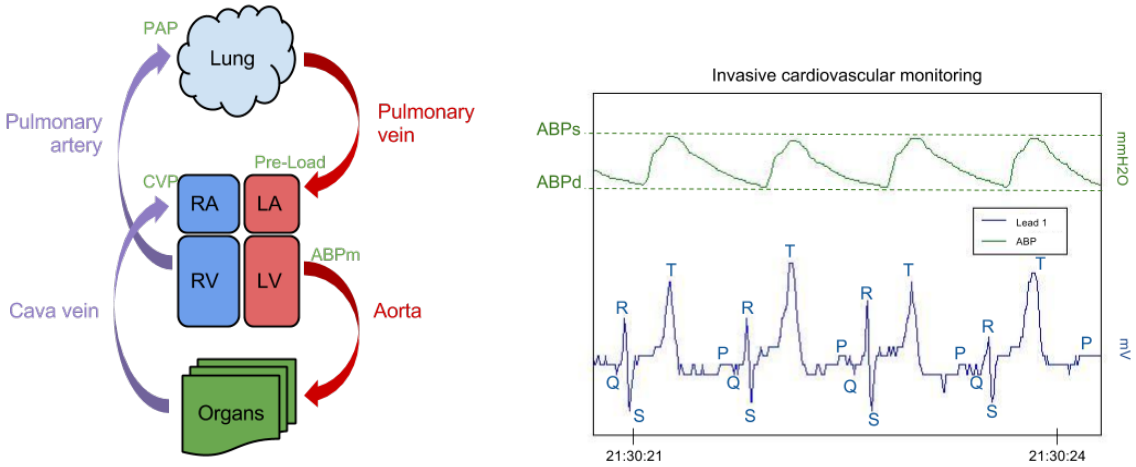


Figure 1.1: *(LEFT) Schematic of cardiovascular system representing the heart (left/right ventricles and atria: LV, RV, LA and RA), arteries, veins, lungs and organs. Sections in red indicate blood with high-oxygen concentration coming from lung and going to organs. Blue sections indicate blood with lower oxygen concentration coming from organs and going to the lungs. In green, the pressure potentially recorded at each heart chamber: Central Venous Pressure (CVP), Pulmonary Arterial Pressure (PAP), Mean Arterial Blood Pressure ABPm and the pre-load. (RIGHT) Four cardiac cycles extracted for a typical ICU patient showing the blood pressure (green) and the ECG (bottom). Systolic and diastolic values are indicated on the ABP waveform while the different cardiac electric waves (P, Q, R, S, and T) are indicated on the cardiac waveform.*

Arteries deliver the blood from the heart to organs, and (apart from the pulmonary artery) carry oxygenated blood ($SaO_2 > 90\%$) under the driving force of a characteristically high blood pressure ($ABP_{mean} > 60$ mmHg);

Veins carry the blood from organs to the heart, and mostly (with the exception of the pulmonary veins) carry low oxygenated blood ($SvO_2 < 80\%$) with high CO_2 levels at low blood pressure (MAP< 20 mmHg).

As seen in Figure 1.1, the heart is composed of four chambers that contract according to a very specific sequence:

1. diastole: the heart is relaxed and blood arriving from the veins is filling; arterial blood pressure at this moment is called Diastolic Arterial Blood Pressure ($ABP_{Diastolic}$) and has typically values of around $60 - 90$ mmHg for healthy adults;

2. left atrial systole: blood arriving from the veins fills the atria with about 40% additional blood volume;

3. ventricular systole: the left ventricle contracts and ejects the blood to the arteries creating a sharp rise in blood pressure to about $95 - 140$ mmHg (ABP_{Sys}).

The electrical activity generated by moving charged particles in the cells during contraction of the cardiac muscle can be recorded non-invasively with electrodes located on the chest. The signal recorded is called the Electrocardiogram (ECG) and is a widely-used diagnosis tool. The sequence of contractions described above generates very typical waveforms denoted as P, Q, R, S, and T that characterise the ECG and usually serve as a basis for the diagnosis of heart conditions. However figure 1.1 provides a good example of how this analysis can be complicated: the T-waves presented in this example have a larger amplitude than the R-wave, an unusual phenomenon. For instance, ventricular contraction generates a large electrical wave (the R wave) that is "easily" identified and can be used to derive RR time series from which Heart Rate (HR, number of heart beats per minute) and can be derived.

Another broadly used monitoring technique consists of inserting a cannula at different locations of the cardiovascular system to monitor pressure near different heart compartments as seen in Figure 1.1. The values of arterial blood pressure at different times of the cardiac cycle have different meanings: $ABP_{Diastolic}$ relates to vascular tone and peripheral resistance, Systolic Arterial Blood Pressure ($ABP_{Systolic}$) relates to the heart contractility. Mean Arterial Blood Pressure (ABP_{Mean}), which is often approximated as $ABP_{Mean} = \frac{1}{3}ABP_{Systolic} + \frac{2}{3}ABP_{Diastolic}$ also provides an aggregated measure of blood pressure and has to be distinguished from the Mean Arterial Pressure (MAP) that is the integrated area under the blood pressure waveform for each beat.

Cardiac Output (CO) measures the blood flow (in litres per minute) from the left ventricle and is directly relates to tissue perfusion, which is of paramount importance for the monitoring of cardiac function during sepsis. CO can be related to Heart Rate (HR) according to the following formula:

$$CO = HR \times SV \tag{1.1}$$

where Stroke Volume (SV) is the volume of blood ejected from the heart during ventricular contraction. Unfortunately, measuring CO cannot be achieved directly and alternative techniques with various degrees of fidelity and invasiveness are used instead. According to Equation 1.1, CO can be controlled by modifying HR or SV, which itself depends on:

- pre-load: the volume of blood available at pre-contraction, which is clinically estimated by using CVP as a proxy for it and can be artifically elevated by rapidly giving the patient some volume of fluid (typically $V > 250$mL) through a venous catheter;
- contractility: the ability of the heart, in terms of power and energy, to achieve fluid exchange between its compartments, which can be estimated from the Pulse Pressure (PP) ($ABP_{Systolic}$-$ABP_{Diastolic}$);
- after-load: vascular peripheral resistance, vaso-dilation, and blood coagulation.

Control of these variables to restore cardiac output and perfusion are key in the management of sepsis and further details of these mechanisms will be provided in section 1.5.2.

1.2.3 The autonomous nervous system

The cardiovascular parameters described in the previous section are controlled by the body in order to maintain adequate tissue perfusion and meet various organs' metabolic demand. Figure 1.2 represents some mechanisms involved during the regulation of haemodynamics: each mechanism is characterised by an amplitude and a specific time scale (temporal location and width of the peak of activity). The mechanism can be largely separated into two groups: in red, rapid-response mechanisms regulated by the Autonomous Nervous System (ANS), and in yellow, longer time scale mechanisms such as activation of kidney function that potentially offer a much larger response.

The ANS is composed of two opposite systems: the sympathetic (Σ) and the parasympathetic ($\overline{\Sigma}$) system, also named "Fight and Fly" and "Rest and Digest", respectively, which

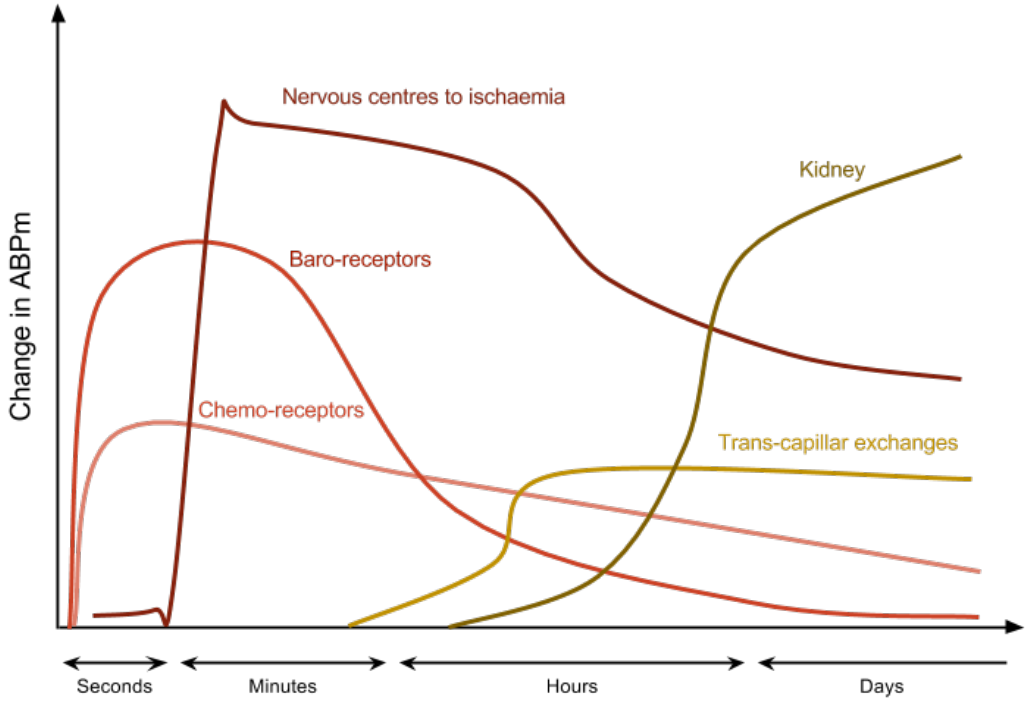


Figure 1.2: *Amplitude of Arterial Blood Pressure (ABP) changes for different regulation mechanisms plotted against time. Rapid responses, such as ANS regulation (in red), have a lower amplitude than longer time-scale mechanisms such as kidney activity and transcapilar exchanges (in yellow).*

illustrates their primary function. Both collect information about the cardiovascular system, integrate it, and trigger an adequate response:

Sensors Baroreceptors are pressure sensitive molecules to be found on the wall of vessels and continuously report on ABP by transduction of pressure into an electric signal. Chemoreceptors on the other hand only trigger signals when abnormal levels of blood gases are detected such as during hypoxia (Partial pressure of O_2, $PaO_2 < 75$ mmHg). These sensing molecules are mostly found in the carotid bulb and the aortic arch, which are connected to the brain stem through cranial nerves IX (Herring nerve) and X (vagal nerve), respectively.

Integration centres Neural information is directed towards the Solitary Track Nucleus (STN), located at the medulla in a small part of the brain stem called the lower pons. This area is divided into two parts: a pressive and a depressive one that relate to Σ and $\overline{\Sigma}$, respectively. The activation of any of these areas is accompanied by the in-

hibition of the other one, which is referred to as the sympathetic/para-sympathetic balance ($\Sigma/\overline{\Sigma}$ balance). Similarly, the presence of hypoxia in the medulla triggers a strong activation of the pressive region and a strong inhibition of the depressive one. Activation of these regions is then translated into both a nervous and hormonal command.

Neural control The heart is connected to both the Σ and $\overline{\Sigma}$ efferent neurons and its activity can therefore be up- and down-regulated by the adjustment of its contractility and frequency. Blood vessels on the other hand are only under $\overline{\Sigma}$ nervous control meaning that only vasodilation can occur via this activation pathway.

Hormonal control Σ and $\overline{\Sigma}$ centres also express their command by the release of norepinephrine and acetylcholine, respectively. These neurotransmitters will circulate in the blood and bind to receptors attached to blood vessels and in the heart. Adrenalin is a product of norepinephrine that generates a strong Σ activation characterized by a sharp rise in HR and ABP_{Mean}.

1.3 Pathophysiology of sepsis

The response of the immune system to the presence of external elements in the body described in section 1.2.1 is a healthy mechanism: local inflammation provokes local vasodilation and increased permeability of arterial walls to ease the delivery of specialized agents to the site of infection. However, following a poorly understood dysfunction (section 1.3.2), this response goes from local to *systemic* threatening the body's ability to maintain homoeostasis (the preservation of biological constants such as body temperature). This translates to Systemic Inflammatory Response Syndrome (SIRS) – see section 1.4 – and early signs of hypotension that are often the first identified symptoms of infection. This occurence is often shortly followed by admission to the ICU. Figure 1.3 presents the different stages of sepsis: the real sequence of events, the time of diagnosis and that of treatment that are all typically mismatched during sepsis. Because SIRS is

also exhibited in non-infectious insults such as burns [38], a blood test would look for a possible infectious agent shortly after admission. Unfortunately, blood cultures can have a long Turnaround Time (TAT) – typically 24 hours – and treatment to fight the infection together with its possible side effects is sometimes initiated prior to actual diagnosis. The following complications of sepsis on the global haemodynamics of the patient leads to Multiple Organ Dysfunction Syndrome (MODS), septic shock and ultimately death as detailed in section 1.3.3.

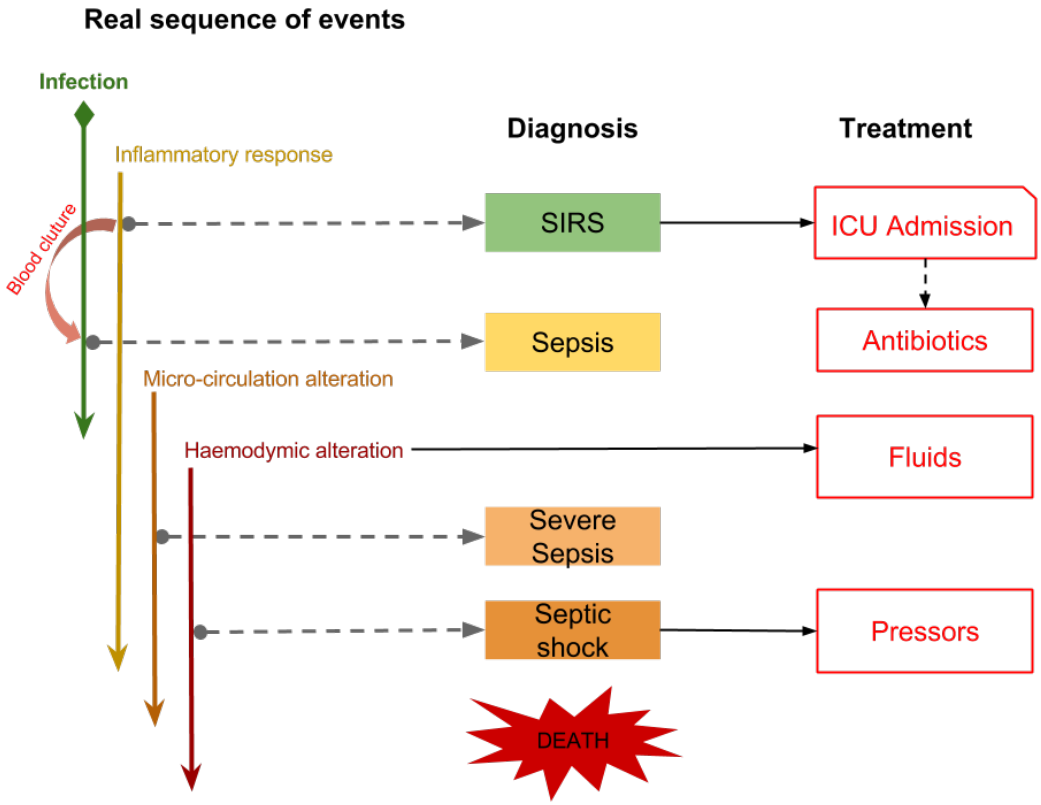


Figure 1.3: *Course of events during the evolution of sepsis. The real sequence of events (downward arrows) is shown on the left and is typically different from the diagnosis sequence (middle). Indeed, infection is asymptomatic before SIRS that often does not declare before a few hours. As a consequence, treatments (red boxes on the right) are often given preventively despite potential side effects.*

1.3.1 Step 1: Recognition of microbial-associated molecular patterns

The mechanism leading to sepsis is usually triggered by the recognition by the innate immune system of very specific internal danger signals [9]. These signals are often com-

posed of cell wall components belonging to two families of bacteria: the Gram-positive and negative bacteria with LipoTeichoic Acid (LTA) and LipoPolySaccharide (LPS), respectively [9]. Likewise, other molecular patterns such as flagellin, lipopeptides, pipoproteins, or PeptidoGlycan (PG), from bacterium or fungus, can also trigger the host response [9, 78, 291]. These pathogen molecular patterns circulating in the blood can connect with two kinds of components: free components or outer-membrane receptors of specialized cells (leukocytes). The LPS animal model [78] studies the bimolecular pathways leading to sepsis after injection of various doses of LPS. In this model, the endotoxin first binds to LipoPolysaccharide-Binding protein (LPB) before presentation to adequate receptors (including the CD-14 protein) that are found in two forms: phagocytes that are surface-bound (mCD14) or soluble in the plasma (sCD14). The LPB-CD14-LPS complex then associates with another protein called Lymphocyte antigen (MD)-2-Toll-Like Receptor (TLR)-4, present on the outer-membrane of leukocytes. This connection finally initiates an internal transduction cascade that results in the synthesis and release of proteins involved in the inflammatory response.

The intracellular signalling cascade involves hundreds of activation and down-regulating factors leading to the transcription of DeoxyriboNucleic Acid (DNA) sequences into proteins such as cytokines. Cytokines are small protein molecules involved in the communication between the different cells and in particular those of the immune system. Table 1.1 lists some cytokines known to have a determinant role in the evolution of sepsis [88, 285, 224] including "acute-phase proteins" that are particularly involved in the early stages of inflammation. This simplified description illustrates a few mechanisms and agents involved during the recognition of LPS, and also gives an overview of the potential complexity of the mechanisms at stake. Pierrakos and Vincent [224] provide a review of molecular biomarkers that have been associated with the severity of sepsis in vivo and in particular the plasma concentration of most proteins listed in Table 1.1.

Table 1.1: *A list of cytokines involved in the inflammatory response and evolution of sepsis. Their role during sepsis can be inflammatory (I) or anti-inflammatory (A). The table shows the main source of production, the type of cells targeted and their main function. Most cytokines would however also be produced by other cells and similarly interact with others, leading to a much broader variety of mechanisms than those described here, yet of smaller amplitude.*

Name	Type	Source	Target	Function
IL-6	I	Macrophages	Neutrophils	Proliferation and differentiation
IL-8	I	Macrophages	Granulocytes	Migration and activation
IL-10	A	Monocytes	Macrophages	Inhibit cytokine production
MIF-1	I	Leukocytes	Macrophages	Regulation of macrophage function
If-γ	A	T-cells	T-cells	Regulation of immune response
TNF	I	Macrophages	Leukocytes	Migration, proliferation, phagocytosis

1.3.2 Step 2: Pro and anti-inflammatory response

As explained in sections 1.2.1 and 1.3.1, the inflammatory response orchestrated by cytokines is balanced by anti-inflammatory mechanisms in order to contain inflammation to levels proportional to the initial insult. There is now long-standing evidence in the medical literature that sepsis is at least partly imputable to a dysfunction of this inflammatory/anti-inflammatory balance [9, 297, 294].

The early phase of inflammation is partly coordinated through the release of Tumour Necrosis Factor (TNF), InterLeukin (IL)-1 and InterFeron (If)-γ. The release of these cytokines in the blood increases Nitric Oxide (NO) blood levels that relaxes smooth muscles and generates vasodilation. Simultaneously, other agents will increase the arterial wall permeability to allow for immune system agents to reach the site of infection in the interstitial milieu. Histamines are released by basophils after their activation and probably also play an important role during sepsis. When the inflammation goes systemic, the increase of cardiovascular volume to be filled together with the dramatic reduction of available fluid (increased permeability of vessels wall) leads to a characteristic drop in blood pressure, which alters micro-haemodynamics and threatens the body's ability to maintain adequate perfusion.

Simultaneously, the anti-inflammatory response tries to maintain inflammatory response in the appropriate range of activity and ensure the stability of the system. This down-regulation of the inflammatory response is supported by the circulation IL-4, 6, 10 and 13, and If-α cytokines that have been identified as playing a central role in this response [17, 87]. Unfortunately, over-activation of the anti-inflammatory response leads to immune depression that favours the further development of the initial infection and creates an opportunity for additional infectious foci. The mechanisms by which this immunosuppression occurs are still active fields of research and certainly the source of much controversy [297].

1.3.3 Step 3: Shock and organ failure

Both the inflammatory and anti-inflammatory processes are healthy when moderate and have beneficial effects leading to successful elimination of infectious agents. In sepsis however, the excessive systemic inflammatory response associated with a prolonged and non-adapted release of cytokines leads to an excessive recruitment of both the innate and adaptive immune system. The exact sequence of events occurring during sepsis is still an area of active research. Such research is difficult outside animal models since symptoms and modification of vital signs occur late in the process leaving little room for clinical observation. The current understanding of mechanisms leading to disturbed haemodynamics involves:

Vasodilation caused by NO, histamines and release of cytokines, which decrease the vascular tone and the systemic vascular resistance;

Increased vascular permeability due to TNF-γ and histamine-1 (both responsible for fluid transfer to interstitial tissue), which when activated reduces blood volume and consequently the pre-load;

Myocardial Infarction is the result of the circulation of myocardial depressant factors such as TNF-α, IL-1β and NO that impair cardiac contractility.

The current understanding is that, during the hypotensive regime, the prolonged hypo-perfusion of tissues leads to apoptosis (programmed cell death), slowly decreasing some organs' ability to function and play their role. The dysfunction of more than one organ is known as MODS, which is naturally associated with the worst outcome, since the activity of every organ is of uttermost importance while the body is fighting a life-threatening external assailant. In addition to this, the induced tissue necrosis reinforces the presence of inflammatory markers meant to activate the cleansing of dead tissues, which further disrupts the inflammatory/anti-inflammatory balance. More recently, Haddad and Harb [113] have underlined the role of Hypoxia-Inducible Factor (HIF) in cytokine regulation and widespread derangement in mitochondrial Adenosine TriPhosphate (ATP) synthesis, suggesting a protective mechanism, that is shutting down metabolism to protect tissue from injury consecutive to ischaemia (a lack of oxygen). In this context, organ failure can be seen as a healthy, and potentially reversible, attempt by the body to protect itself from the lack of metabolites. This could explain why *post-mortem* observations of tissue have not found signs of necrosis and apoptosis spread enough to explain organ failure as noticed by Hotchkiss et al. [136].

1.4 Clinical definitions of sepsis

1.4.1 Early recognition of infection and complication

Interestingly, an early reference to sepsis in the modern medical literature [301], states the problem in words that sound surprisingly modern, while distinguishing the initial insult to the host response:

> In considering the aetiology of post-operative sepsis and other hospital infections, the bacteriologist is apt to think in bacteriological terms and to seek to explain the incidence of infection by reference to the sources from which the patient could have become infected. The clinician, on the other hand, is inclined to think of clinical reasons why some patients become septic and

Table 1.2: *SIRS is defined by the presence of at least two of the following conditions. A patient can be considered septic if a documented source of infection can be added to the presence of SIRS*

Parameter	Values
Temperature	$< 36°$C or > 38
Heart Rate	> 90 bpm
Respiratory Rate	> 20 bpm or $PaCO_2 < 32$ mmHg
White cell blood count	$> 12.10^9$/l or $< 4.10^9$/l or > 10% immature bond forms

others not. A true understanding of the aetiology of septic complications in hospital patients can be achieved only by a combination of bacteriological and clinical analysis.

Recognition of sepsis as a clinical identity really arose in the 70's with the identification of bacterial shock. Initially, Gram-negative bacteria were identified as being a primary factor associated with complication of infection Shubin and Weil [257], Bone et al. [39] leading to shock. Furthermore in both studies, the use of catheters (venous and urinary) were associated with the infection. Finally, Bone et al. [39] introduced the first elements describing sepsis syndrome with systemic inflammation and organ dysfunction in the presence of infection, ultimately followed by shock.

1.4.2 The 1991 consensus conference

A consensus conference was held in 1991 under the chairmanship of Bone et al. [38]. It proposed to abandon terms such as sepsis syndrome, septicaemia and blood poisoning and promoted instead the use of a newly defined concept of SIRS that is detailed in table 1.2. This term introduces the concept of a host response that is independent of the injury and that may or may not be related to an infection. As a consequence, this definition has lacked specificity: patients with burns for example will also exhibit SIRS. Sepsis syndrome was therefore defined as SIRS plus a documented source of infection.

1.4.3 The 2001 definition

Following the 1991 Consensus conference, additional drawbacks of the proposed definition were identified:

- Over-sensitivity of the SIRS criterion;
- Over-specificity of the definition of infection;
- Lack of overall adoption of the definition of sepsis in the literature;
- New understanding of pathophysiology sepsis;
- Proven needs for early diagnosis and treatments.

As a result, the definition was only partly accepted by the research community and needed improvement. In particular, the outcome of sepsis was known to vary depending on the severity of the clinical observation. Therefore, three additional steps were defined for the course of infection and sepsis [180, 190]:

Severe Sepsis Sepsis associated with one of the following conditions: hypo-perfusion, hypotension and/or organ dysfunction;

Septic Shock Sepsis-induced hypotension (defined by $\text{ABP}_{\text{Systolic}} < 90\text{mmHg}$) despite adequate fluid resuscitation;

Multiple Organ Dysfunction Syndrome (MODS) dysfunction in one or more organs.

The new definition includes a wider range of symptoms, acknowledging the complexity of the condition and giving the clinician larger control of the diagnosis, defining as "septic" a patient matching *some* of the criteria, together with flexible decision boundaries. If the idea of a patient *looking* septic does not constitute a reproducible criterion, it does match bedside reality. Moreover this consensus conference introduced the new concept of stratification in risks with a system called *PIRO* based on Predisposing conditions, the nature and extent of the Insult, the nature and amplitude of the host Response and the degree of concomitant Organ dysfunction as reported by Levy et al. [180].

1.4.4 Latest definitions

The latest definition presented by Dellinger et al. [76] defines sepsis as the presence of *probable* or documented infection together with *some* signs of systemic manifestation of the infection such as fever, tachycardia, positive fluid balance, hyperglycaemia, or abnormal WBC. Severe sepsis is then defined as sepsis-induced organ dysfunction or tissue hypo-perfusion, that are themselves defined from thresholds on lactate levels, urine output, creatinine, or bilirubin. An important distinction is made between thresholds for definitions and targets for treatment. For instance, sepsis induced hypotension is defined as $ABP_{Systolic} < 90$mmHg or $MAP < 70$mmHg, while fluid therapy (presented below) will typically target a MAP of 65mmHg.

In an even more recent article, Vincent et al. [290] reminds us that up to 90% of ICU patients meet SIRS criteria and concludes that any infected patient therefore could be defined as a septic patient. Additionally, similarities shown in response following invasive infection and sterile tissue necrosis, ssuggest that a radical change should occur in the definition of sepsis. In other words, some degree of organ dysfunction should be included in the definition of sepsis [290].

1.4.5 Use of the definition of sepsis

Although sepsis is a long-standing medical condition and widespread in modern healthcare systems, its definition has been changing significantly in the last decade and remains unclear. Yet, the growing understanding of the underlying pathophysiology allowed for the diagnosis of sepsis to become more and more specific [289] even though it is still arguable that there is no gold standard for the definition of sepsis to date [224].

The main problem with the early definitions [38] was that it was too sensitive and not specific enough. Newer definitions [190, 75, 76] are certainly more specific but lack consistency: in particular, the clinical definition contains a non-negligible number of terms lacking precision such as *some*, *probable*, *suspected*, *presumed*, *looking*, *substantial*.

The evolution of the definition of sepsis did not have a major impact on clinical prac-

tice since recommended therapeutic targets are different from definition thresholds [76]. However, a consistent definition would be of special interest for the design of clinical trials and epidemiological studies as detailed in section 1.6. In particular, the absence of a common definition of sepsis hinders accurate comparison between results for different studies. Additionally, as we will further develop in section 2.2.1, the absence of a clear definition for the identification of the population of interest from clinical databases dramatically complicates retrospective studies such as those presented in this items.

1.5 Management of sepsis, severe sepsis and septic shock

As for its definitions, the guidelines for managing sepsis and severe sepsis have seen great changes over the past ten years. These changes were naturally driven by the increasing clinical literature on the topic (see Figure 1). Until 2004, clinicians had to mine this literature and adapt decisions taken at the bedside based on the quality of the evidence presented to them: sample size, entry criteria, and type of study design.

Since 2004, the available evidence has systematically been reviewed by a large panel of leading international experts ($n = 68$ from 30 institutions) in a consortium named the Surviving Sepsis Campaign (SSC) [74, 259]. The committee was asked to look at different treatment options presented in the literature and to follow the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system to assess quality of evidence, taking into account factors such as the design of the study – Randomised Control Trial (RCT) or observational, the quality of the implementation, the population (size, inclusion criteria), and the effect (amplitude and presence of gradient with dose). The recommendations have been updated regularly since then [75, 179, 76].

The guidelines address key aspects of sepsis management: diagnosis, source control, and haemodynamic support. The general view is that early action prevents the worsening of the patient's state: control of the infectious agent prevents further development of the inflammatory response, resuscitation and haemodynamic support restore oxygen supply

in order to prevent tissue necrosis and HIF [113]. Additional supportive therapy targets specific vital functions such as cardiac and renal function to prevent additional insults.

1.5.1 Source identification and control

1.5.1.1 Infection prevention and control

Before any therapeutic consideration, the SSC recommends careful control of potential sources of infections: necrotic soft tissues, peritonitis, intestinal infarction and promotes minimal surgical intervention with the least physiological insult to handle them.

Hospital-acquired sepsis tends to be associated with significantly higher costs than those for which patients may be initially admitted [45]. This certainly reflects the personal burden supported by patients and their families. Hospital acquired sepsis is often the side effect of an invasive intervention assumed to balance risks and benefits to balance risks and benefits positively: elective surgery in the first place, but also cannulation for vascular access, urinary catheters, drains, and mechanical ventilation. As a consequence, it is recommended that these devices be replaced regularly and their use discontinued as soon as the patient's condition allows. Oral and digestive decontamination are simple and effective ways to reduce the incidence of sepsis and severe sepsis in patients under mechanical ventilation.

Check-lists were not mentioned in the latest SSC recommendations despite their great potential in reducing infection [126, 100] and the recent interest of the critical care community [302, 299]. On-going clinical trials may promote broader adoption of such an approach that could become a standard of care in the future [171].

1.5.1.2 Screening and diagnosis

A specific and early diagnosis is of paramount importance in the management of sepsis. Indeed, lack of specificity would result in an overprescription of antibiotic treatments, leaving room for the development of opportunistic resistant strains of bacteria; on the

contrary, a delayed response to the infection allows time for development of the threatening agent, possibly enhancing the already significant inflammatory response. Routine screening of potentially infected seriously ill patients has been associated with decreased sepsis-related mortality [179].

One of the main issues associated with the diagnosis of sepsis is the difficult identification of infectious agents [180, 190]. Good practice for the diagnosis of infection therefore suggests multiple sampling strategies (aerobic / anaerobic and intra-vascular / percutaneously) in addition to timely imaging studies. The antimicrobial therapy (described in the following section) should ideally be initiated after these cultures if it does not induce significant delays (> 45min).

1.5.1.3 Antimicrobial therapy

In an effort to control the source of infection, anti-microbial therapies can effectively support surgical procedures to eliminate infectious agents and limit the extent of the inflammatory response. The goal of therapy as suggested by the evidence reviewed by the SSC committee is to initiate such a treatment within an hour of recognition of severe sepsis or septic shock [76]. One or more drugs targeting *likely* pathogens could be used in adequate concentration on the tissues *presumed* to be the source of sepsis. Antimicrobial therapy (described in the following section) should be re-assessed daily for potential de-escalation to reduce toxicity, costs and development of resistance. Biomarkers such as low levels of procalcitonin (a small molecule specifically synthesized during inflammation induced by bacterial infection) could assist the clinicians in the decision of discontinuation. In any case, the duration of treatment should not exceed a week, except for special cases.

1.5.2 Haemodynamic support

The drop in blood pressure induced by systemic vasoplegia (vasodilation of blood vessels) and increased wall permeability threatens the body's ability to deliver oxygen and

therefore meet the metabolic demand, which increases with the activation of defence mechanisms during the fight against infection. Similarly, bypass products of metabolism are not evacuated and accumulate in cells, disturbing the ionic balance and potentially reaching toxic levels. Inadequate perfusion can lead to tissue necrosis and organ dysfunction, simultaneously increasing inflammatory sources and damaging the body's ability to fight infection. Timely fluid resuscitation can adequately restore organ perfusion and hopefully prevent or reverse dysfunction.

The most significant improvement in the management of sepsis in recent years lies in the Early-Goal Directed Therapy (EGDT) introduced by Rivers et al., which recommended early target numbers for blood pressure and oxygen saturation [243]. Despite the strong evidence provided in the original study, it took a few years before findings got confirmed and broadly accepted by the critical care community [244, 238, 286]. Therefore an initial precisely quantified resuscitation of patients with sepsis-induced tissue hypo-perfusion persisting after fluid-challenge is recommended by the SSC [76]. The resuscitation should not be delayed by ICU admission, which is of paramount importance for data analysis. Precisely, it means that data collected during the ICU stay does not necessarily capture all treatments given to patients like fluid challenges (rapidly giving the patient fluid (typically $V > 250$mL) through a venous catheter). During the first 6 hours of resuscitation, the goal set by the SSC is :

1. CVP between $8 - 12$mmHg
2. MAP above 65mmHg
3. Urine Output (UO) greater than 0.5mL.kg.hr
4. Vena Cava Oxygenation Saturation ($ScvO_2$) greater than 70%
5. Mixed Venous Oxygen Saturation (SvO_2) greater than 65%
6. normalize lactates for patients with elevated lactate as a marker of tissue perfusion

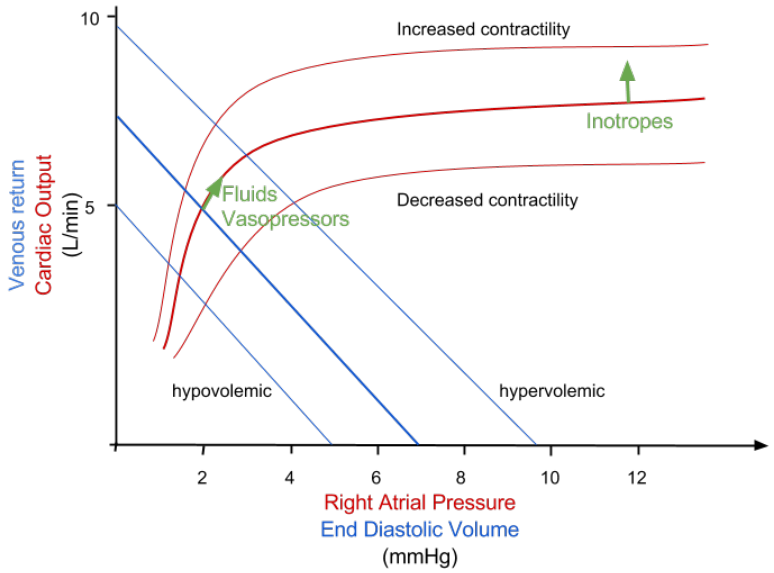


Figure 1.4: *The cardiac function in red is represented by the evolution of the pre-load volume as a function of the right atrial pressure for different contractility. The vascular function in blue is represented by the evolution of venous return as a function of end-diastolic volume for different blood volumes. The cardiac parameters are taken at the intersection of the two lines. In green, the arrows indicate the different treatments that can influence pre-load and cardiac output. This diagram was inspired from [277].*

Unfortunately, an adequate tissue perfusion cannot always be restored in the initial phases of sepsis. In this case, while the source of infection is sought and possibly treated, haemodynamics should be supported to maintain organ perfusion and prevent further organ dysfunction. The mechanisms by which CO is controlled are presented in Figure 1.4. This also illustrates some procedures that can be initiated to maintain targeted blood pressure:

- Fluid therapy to treat hypovolemia;
- Pressors to address systemic vasoplegia;
- Inotropic therapy to regulate inflammatory response.

1.5.2.1 Fluid therapy

The goal of fluid therapy is to prevent hypovolemia. Indeed, as explained in section 1.3.3, the combined effect of increased wall permeability and vasoplegia leads to a decreased

cardiac output and inadequate organ perfusion. The use of fluid resuscitation to fill the patients' cardiovascular system is a first means to restore targeted blood pressure. Different types of fluids have been historically considered for fluid resuscitation:

Crystalloids are aqueous solution of mineral salts , which comprise :

- Saline solutions such as 0.9% Sodium Chloride solution that are isotonic (equivalent concentration to that of the blood);
- Lactated Ringer's (or Ringer's lactate) are slightly hypotonic (smaller NaCl concentration than blood) and are designed to account for the expected patient acidosis. In the UK, Hartmann's solution is the equivalent;
- Dextrose is a sugar solution (glucose) that also provides nutrients;

Colloids are solutions made of much larger molecules such as starch and albumin, preferred because resuscitation is not affected by the increased arterial wall permeability since larger molecules tend to remain in the intra-vascular space [169].

The SSC recommends crystalloids as an initial choice for fluid resuscitation and suggests albumin as a preferred colloid for patients with septic shock for whom *substantial* amounts of crystalloids are required. Use of Hydroxyethyl starchs (HESs) has been discouraged. These recommendations are based on recent RCTs in different regions of the world [110, 213, 223]. The initial recommended fluid challenge for sepsis-induced hypo-perfusion and *suspected* hypovolemia is to achieve a minimum of 30mL/kg, to be continued as long as improvement can be monitored (arterial pressure, heart rate, pulse pressure, stroke volume).

1.5.2.2 Vasopressors

Vasopressor agents are drugs causing constriction of blood vessels. They also refer more generally to agents raising blood pressure, which also include inotropes described in section 1.5.2.3. Catecholamine (epinephrine, norepinephrine, dopamine and phenylephrine)

is a common type of vasopressor that is also called sympatho-mimetic since it exhibits similar behaviour to that of molecules synthesised by the sympathetic nervous system (the *fight and fly* response). Their activity differs slightly as the receptor they target may be found at different surfacic concentration in blood vessels from different organs. Additionally, they affect the cardiac function in various degrees: dopamine for instance has been associated with increased occurrences of tachycardia [235]. Finally, vasopressin is an anti-diuretic hormone that retains blood fluids and also constricts vessels.

Based on the available literature, the SSC recommends to initiate vasopressor therapy to initially target a MAP of 65mmHg, even though targeting macrocirculatory endpoints is regularly challenged [81]. The preferred agent is norepinephrine, to which epinephrine can be added. Vasopressin (up to 0.03U/min) should not be identified as an initial choice of vasopressor but could subsequently be used in addition to others to raise blood pressure further or to replace norepinephrine. Finally, the use of dopamine and phenylephrine should only relate to highly selected patients such as those at risk of tachyarrhythmias.

1.5.2.3 Inotropic therapy

Inotropic agents modulate the force of muscular contractions and are commonly used to regulate cardiac function. They can have both positive and negative effects. Inotropes have different types of actions, the most common working by increasing the level of calcium in the cytoplasm of muscle tissue. Catecholamines and insulin have an inotropic effect. Therefore, the recommendation of the SSC includes the use of dobutamine infusion (20μg/kg/min) in the presence of myocardial dysfunction or signs of hypo-perfusion despite reaching adequate MAP [99, 125, 76].

Negative inotropic agents have also surprisingly been investigated for the treatment of septic shock. At first, it certainly seems contradictory to give negative inotropic agents to patients for whom cardiac function is compromised. The underlying hypothesis is that tachycardia drives the heart in working areas where filling time is not sufficient: a

decreased heart rate therefore increases the filling-time and is not associated with a drop in SV while preserving the cardiac muscle functioning at an elevated level, hence preventing muscle fatigue; unfortunately this has only been investigated on large animals [2]. Other metabolic and immunomodular effects of β-blockers seem to play in favour of its use as a treatment for sepsis [211]. A small retrospective study ($n = 88$) could not relate the use of β-blockers to a better outcome in patients with sepsis [111].

1.5.2.4 Corticosteroids

The use of corticosteroids for treatment of severe infection was introduced more than half a century ago [114] and is broadly used today amongst physicians [76, 75, 179, 74, 75]. There is still however an important controversy between experts about the benefit-to-risk ratio, which comes down to targeting the right population of patients for this treatment alternative [13]: "Who", "When" and "What" ?

We presented in section 1.3.2 the strong evidence showing that uncontrolled systemic inflammation plays a central role in severe sepsis. Coincidentally, glucocorticosteroids' molecular mechanisms seem to perfectly fit this mechanism via fast non-genomic (decreased platelet aggregation [200]) and genomic effects promoting the anti-inflammatory response after a few days of exposure [281]. Other important effects related to the restoration of cardiovascular homoeostasis have been reported even though only partly understood [13]. In particular, mean arterial pressure was restored in septic shock patients treated with norepinephrine and the effect was found to be strongly correlated to the pressor dose response [10]. Recent evidence [76, 13] suggests that only patients with septic shock and specific response to vasopressors therapy may benefit from corticosteroid therapy. Precisely, a systematic review of high quality RCT trials showed a non-significant trend towards reduction of mortality in the most severe subgroup of patients [220].

1.5.3 Other supportive therapy

Naturally, there are many other aspects in the management of septic shock that cover support of failing organs:

Blood products There is a general movement towards less use of blood products in severe sepsis patients with haemoglobin targets of 7 g/dL, which is motivated by a weak rationale in this population together with elevated risks of complication;

Mechanical ventilation is a key component of the management of Acute Respiratory Distress Syndrome (ARDS) and, as with for most therapy, discontinuation should be sought as early as possible to stimulate respiratory function and prevent risks of mechanical ventilation associated pneumonia;

Sedation and particularly neuroblockade agents should also be minimized and discontinued as soon as possible as prolonged use has been associated with a poor outcome;

Glucose control should be initiated with insulin therapy as soon as two consecutive measurements above 180 mg/dL are observed;

Renal Replacement Therapy (RTT) can overtake failing renal function and continuous techniques (instead of intermittent haemodialysis) should be preferred when available;

Prophylaxis of deep vein thrombosis should be administered daily to severe septic patients preferably using low dose heparin;

Nutrition Oral or enteral feeding is preferred to fasting or provision of only intravenous glucose (dextrose solutions) and should be administered to up to 500 kcal per day (low-dose feeding).

Last but not least, the SSC committee recommends that the goals of care are carefully communicated to patients and their families no later than 72 hours following admission to the ICU.

1.6 Epidemiology and cost of sepsis

1.6.1 Epidemiology of sepsis

As we have seen in section 1.4, the identification of patients matching the definition of sepsis from any database is difficult: the definition better fits bedside reality than that found in clinical databases. In particular, the documentation of infection lacks both sensitivity (presence of false negatives) and specificity (false positives). Altogether, this leaves room for misdiagnosis and results in high variance in epidemiological studies. In addition to this, variations in populations (rules for admission to the ICU) and local practices between healthcare systems can amplify further differences observed in survival rates. For instance, Parrillo et al. [218] report an annual number of cases of sepsis in the USA of $400,000$, out of which $200,000$ develop into severe sepsis (50%) and of which about $100,000$ die (25%). Ten years later, Angus et al. [6] carried out a wide study on severe sepsis in the US ($n = 6,621,559$) indicating a greater than two-fold increase in incidence ($751,000$ cases per annum). This increase almost certainly relates to a change in definition, rather than a real increase in incidence, since the equivalent figure of half a million cases a year for the United States of America (USA) are also reported by Martin et al. [189] over the same period of time.

Yet, these obvious limitations can be circumvented in order to extract trends and the increase of incidence still remains a widely accepted fact. For instance, McGowan et al. [194] report a ten-fold increase between 1930 and 1960 (about 30% per year). Later studies however suggest that the increase in incidence of sepsis may have reached a plateau around $1-2$% of hospital admissions [101, 6]. The difference observed between studies are thought to originate from the evolution of both the definition of sepsis and diagnosis techniques. After correcting for the evolution of the American population, Martin et al. [189] still found an annualized increase of 8.7% in incidence. According to Dhainaut et al. [78], the reasons for such an increase are:

- the spread of wide spectrum antibiotics that allowed the emergence of a resistant

strain of bacteria;

- the increase of invasive procedures and therapeutic devices;
- an ageing population (sepsis incidence increases with age);
- the improvement of diagnostic capabilities.

Other interesting figures include a slight predominance of sepsis in males of around 10% [78, 6], in non-white populations [189], and a ten-fold increase with age from 0.2 per thousand in infants to 26.2 per thousand in the elderly. Gender is also an independent predictor of mortality during sepsis, which could possibly reflect the role of sex hormones in immunomodulation [142].

Figures on the mortality rates of sepsis are relatively consistent if compared to the numbers reported for incidence. In Europe, severe sepsis has a mortality rate ranking from 27% to 54% for sepsis and septic shock, respectively [292]. In the US, severe sepsis patients have a similar mortality rate of 28%, which is highly age-dependent, from 10% in infants to 38.6% in elderly and increasing to half of patients experiencing shock [6].

1.6.2 Costs of sepsis

The costs of sepsis have been evaluated to be between $\$22,000$ and $\$50,000$ per patient visit, resulting in a total cost of $\$16.7$m per annum in the US [230, 6, 189, 305]. While mortality related to sepsis is decreasing steadily, the incidence and associated costs are increasing. In addition to this, there is recent evidence that the long-term complications in patients who recovered from sepsis, which were referred to as *lingering consequences* [297] such as cognition impairment and immunosuppression, could have a disastrous impact on public health [8, 7]. This effect, so far poorly explored, could gain increased interest amongst researchers with the emergence of Electronic Medical Records (EMRs) easing the follow-up of patients.

1.7 Estimation of sepsis severity

A number of scoring systems have been developed for use with critically ill patients to determine disease severity and predict mortality. Commonly used outcome prediction scores including the Acute Physiology and Chronic Health Evaluation (APACHE) scores [313], Simplified Acute Physiology Scores (SAPS) [97], and the Mortality Probability Models (MPM) [131]. These scores, described in further detail in section 3.3.1, are usually based on a combination of variables that reflects pre-existing conditions, baseline admission status and physiologic derangement due to acute illness.

The uses of such prediction models mostly depend on the accuracy that can be achieved. Even though prediction in the ICU allows correct outcome identification in about eight patients in ten, their use at the individual level is not recommended [293]. In fact, these models are applied for research purposes at the population level for the benchmarking of ICU performance, comparison of variation in practice between healthcare systems [50, 312, 155], and could ultimately be used for triage in the ICU [311]. Application of severity scores in the sepsis population includes: (i) comparison of population severity between clinical trials [298, 37, 147]; (ii) evaluation of additional predictive power of new biomarkers of severity [207, 237, 236]; (iii) ultimately, triage [254] and risk stratification application for the sepsis population [49, 147].

Despite their importance for clinical research on sepsis, the performance of general ICU physiology-based scores under-performs for the sepsis population [18, 70]. This is partly due to case-mix variation (higher severity for this population for this population and the importance of other predictors), which stresses the need for customization [202, 313]. Furthermore, additional improvement of performance for these models of severity could allow for optimal risk stratification, allocation of resources, and ultimately an adjustment of the care plan to better balance risks and benefits. Additionally identification of markers of severity could offer new insights into the pathophysiology of sepsis, suggesting potentially new therapeutic approaches [161, p. 254].

A fundamental aspect of the work we propose here is the study of patients' physiology

variations over time and in particular during specific events such as hypotension as well as in response to treatments like fluid challenge and use of pressors. The underlying hypothesis is that the specific response to endogenous (hypotension) and/or exogenous (treatment) impulses reflects a dimension of the patient's physiology that could not be captured with static measures. The population of patients with severe sepsis is suited to this hypothesis in particular, as hypotension is present by definition and treatments are normalised to follow the guidelines introduced earlier in this chapter.

Conclusion

The history of sepsis goes back to the great ancient civilisations and illustrates how the lack of rational thinking can mislead medical practice for hundreds of years. This strongly supports the systematic and rigorous analysis of available data to strengthen evidence. Today, the understanding of the aetiology of sepsis reveals a hugely complex pathophysiology involving all the constituents of the body: uncontrolled inflammatory response set up by the immune system threatens the vital homoeostases usually maintained by the cardiovascular system and the ANS. Yet, lack of understanding of the underlying physiological mechanisms has weakened the consistency of definitions and guidelines for many years and only recently have these converged. As a consequence, definitions have become more specific and treatments have changed to steadily reduce the mortality associated with the different stages of the condition. Yet, the impact and cost associated with sepsis, together with the increase in incidence and high mortality rate, constitute a strong rationale for improving knowledge about the condition. In particular, severity scores have proven to be very useful for research in the field of sepsis and could even translate to clinical practice given some substantial improvements. The aim of this thesis is therefore to improve the state-of-the-art estimation of severity in a population of severe sepsis patients.

Chapter 2

Study population and data extracted

2.1 The MIMIC-II database

2.1.1 An introduction the MIMIC-II database

The traditional paradigm of modern science consists of the acquisition of data according to a protocol designed to test a hypothesis; conclusions can then be drawn after adequate analysis of the data. Because of the tedious and costly process of data acquisition, datasets are often kept for internal use and no external validation nor comparison can be achieved easily. The MIT-BIH arrhythmia database [197, 198] was probably one of the first examples of a freely available database, finally allowing direct comparison of techniques for ECG-based detection of arrhythmias. Shortly after, the *Physionet* project was created, offering a collection of datasets and tools dedicated to the study of physiology with a focus on the analysis of vital signs [107].

Beyond the need to replicate studies and have common datasets to compare performance, clinical databases offer much wider possibilities. It was recently reported that as much as $\$750$b is wasted yearly in the US healthcare system [35], while ICU spending accounts for about 1% of Gross Domestic Product (GDP) [153, 116] and about 60% of that amount in other western countries [93]. Despite the large quantities of cash injected into these healthcare systems, it is still commonly accepted that only between 10 to 20%

of decisions are based on evidence [109]. Professor Archibald Leman Cochrane who famously promoted effectiveness in healthcare [64] and initiated the Cochrane Foundation believed this number to be below 10%. However, RCTs - the pillar of evidence acquisition - have shown important limitations including excessive costs and delays, together with a poor efficacy: one in seven RCTs did not show a clear benefit for patients despite previous evidence [214]. The ICU environment is characterised by large amounts of heterogeneous data and the recent emergence of EMRs paves the way for automated collection and analysis of large datasets. Altogether, these elements built a strong case for the use of clinical databases and data mining techniques as a new paradigm for generation of evidence in critical care.

The Multi-parameter Intelligent Monitoring in the Intensive Care II (MIMIC-II) database is the result of a collaborative project between the Laboratory of Computational Physiology (Harvard-MIT, Health Science and Technology), Philips Medical Systems and Beth Israel Deaconess Medical Centre (BIDMC) aiming at the development and evaluation of advanced ICU monitoring systems [248]. Patient information was collected from different data sources primarily the BIDMC Hospital Information System (HIS) and social security records and merged. To date, the database (version 2.6) contains data from 26,655 patients collected over 33,361 ICU stays from 2001 to 2007 for whom thousands of variables are collected.

2.1.2 Dealing with de-identified medical data

In order to release the database publicly, records were de-identified with approval of the BIDMC Institutional Review Board (IRB) [248]. Patterns matching bespoke dictionaries were removed from free texts discarding ZIP codes, phone numbers, names, and places. All dates (starting with hospital admission) were randomly shuffled to years 2,000 to 4,000 while preserving the exact sequence of events. For instance, the time from hospital admission to admission to the ICU (also called pre-ICU length-of-stay) should be consistent. Similarly, the day of the week, time of day and period of the year at the time

of admission are also consistent to allow studies on variation in practice at different times of the year, week, and day. Finally to match patient information across different sources, while preserving privacy, each patient was allocated random subject, hospital-stay, and ICU-stay unique IDs. For instance, demographic data, administrative billing codes, and bedside monitor data are stored in relation to a subject, hospital and an ICU unique ID.

2.1.3 Database architecture

HISs have grown in an organic manner since computerization of record systems [122] to serve administrative, economic and medical purposes [123, p1-11]. As a consequence, patients' data is spread over independent databases, which can be nation-wide, hospital-wide or designed for an ICU service. These independent sources of information were mapped to different tables of the MIMIC-II database [248]. The database was thereby naturally designed as a relational database (Oracle 11g), consisting of different tables (or *relations*) composed of rows (*tuples*) and columns (*attributes*). A row typically describes an entry in the database (for instance a blood sugar level result) with several columns defining the record: the unit, the time of recording and usually also at least one *foreign key* (the different unique IDs) that connects to other tables in the database. For instance, Table 2.1 describes the attributes for a lab result stored in the *LABEVENTS* table and shows how it relates to a specific subject, hospital, ICU stays. It is also a table indicating what blood level is described by the record. The language used to manipulate the data in a relational database is the Structured Query Language (SQL) [260], which was used in this work to identify the cohort and extract the data.

A consequence of the architecture described here is that the same information with a slightly different meaning can be extracted from different tables. For instance, lab results from blood withdrawals will show in the *LABEVENTS* table (extracted from the laboratory information system and described in Table 2.1) as well as in the *CHARTEVENTS* (information filled in by nurses on bedside monitors). The values are prone to different types of noise and may have different time stamps. The correct extraction of the right

Table 2.1: *Description of the* LABEVENTS *table in the MIMIC-II database showing different columns including* foreign keys *linking to other tables of the database (indicated with *)*

Name	Comment
subject identification number (*SUBJECT_ID*)*	The unique patient identifier
hospital stay identification number (*HADM_ID*)*	The hospital admission
ICU stay identification number (*ICUSTAY_ID*)*	The ICU stay id
ITEMID*	The identifier for the laboratory test name
CHARTTIME	The date and time that the test relates to
VALUE	The result value of the laboratory test
VALUENUM	The numeric representation of the laboratory test if the result was numeric
FLAG	Flag or annotation on the lab result
VALUEUOM	The units of measure

data from the MIMIC-II database requires skill and expertise. Yet the database, thanks to the number of patients available, the variety and granularity of data provided, remains a unique tool for the exploration of the severity of disease.

2.1.4 Type of patients and services

The terminology of critical care greatly varies from country to country and even between hospitals. Admission rules mostly define severity and the range of patients encountered in these services. Generally speaking, patients admitted to ICUs require closer monitoring of vital functions or the use of specific procedures. It is not only patients presented with life-threatening conditions that meet these criteria. All patients with trauma, cardiac surgery, organ dysfunction, infection or haemodynamic instability can be encountered in the ICU representing a broad range of severity. Table 2.2 describes the populations of patients admitted to the different ICU services at BIDMC, primarily the Medical ICU (MICU). Interestingly, in-hospital mortality varies from 8.4% in Cardiac Surgery Recovery Unit (CSRU) to 14.5% in MICU, while CSRU patients presented with the most severe indicators at admission with SAPS-I of 17 $(14-20)$ and Sepsis-related Organ Failure Assessment (SOFA) of 8 $(6-10)$ against 13 $(9-17)$ and 4 $(2-7)$ for the MICU patients (the reader can refer to section 3.2.1 for a description of these severity indicators). This could reflect the important physiological insult experienced by cardiac-surgery patients who mostly recover from it. No patient from the Neonatal ICU (NICU) was considered for this work as their physiology differs too much from that of adult patients.

Table 2.2: *Description of populations in different ICU types: Coronary Care Unit (CCU), CSRU, MICU, Surgical ICU (SICU), and NICU. It shows the number of patients in the database (v2.6), ICU mortality rate, Length of Stay (LOS), and median, (25^{th} - 75^{th}) percentiles for age, height, weight, severity (SAPS-I) and organ dysfunction (SOFA) scores. Refer to section 3.2.1 for a description of SAPS-I and SOFA.*

Service type	Count	Mortality Hosp. (ICU)	Length of stay ICU (days)	Age (yr)	Height (cm)	Weight (kg)	SAPS-I (-)	SOFA (-)
NICU	8,080	0.6 (0.6)	0.76 (0.11-9.11)	N.A.	44.5 (41.0-48.0)	2.09 (1.26-2.87)	N.A.	N.A.
MICU	13,258	14.5 (9.0)	2.09 (1.09-4.18)	64.0 (49.5-78.0)	167.6 (160.0-177.8)	75.3 (63.0-90.1)	13 (9-17)	4 (2-7)
SICU	8,091	11.5 (7.8)	2.39 (1.23-5.23)	61.4 (46.8-76.2)	170.2 (162.6-177.8)	77.0 (65.0-90.0)	13 (9-17)	4 (2-7)
CCU	4,854	9.1 (6.6)	2.00 (1.02-3.94)	71.4 (59.1-80.7)	170.2 (162.6-177.8)	79.0 (66.0-93.0)	11 (8-15)	3 (1-6)
CSRU	6,139	8.4 (2.6)	2.10 (1.13-3.95)	67.2 (57.0-76.2)	170.2 (162.6-177.8)	80.2 (68.6-93.3)	17 (14-20)	8 (6-10)
TOTAL (Adult)	32,346	10.9 (7.1)	2.12 (1.12-4.25)	65.5 (51.7-77.7)	170.2 (162.6-177.8)	77.3 (65.0-91.1)	13 (10-17)	5 (2-8)

2.1.5 Description of available data

The number of variables in the MIMIC-II database is vast: about five thousand are available in the chart events table and more than seven hundred in the laboratory events table. Many of these variables have similar meaning and could be merged while others are present in too few patients to be useful. Yet, the amount of available covariates constitutes a typical challenge of clinical data-mining as we will further develop in chapter 5. Information available in MIMIC-II generally includes :

Demographics age, gender, admission type and source, ethnicity, religion and, social security group (at admission);

Laboratory results results from all the blood samples (a few times a day);

Vital signs nurse-verified values entered in bedside monitors including HR, ABP, Arterial Oxygen Saturation (SpO_2), and Breathing Rate (BR) – approximately sampled once an hour;

Medication given to patients and mode of administration;

Administrative data used for billing and administrative monitoring (at discharge);

Free text notes that comprise nurse notes, radiology reports and discharge summary;

Outcome In-hospital mortality, in-ICU mortality and date of death (outside the hospital) when applicable.

2.1.6 Use of ontologies in the database

Ontologies are structured dictionaries designed to provide a relational and computer-friendly representation of a domain. Medical ontologies have been developed to meet application-specific requirements, mainly for secondary use of the clinical data serving administrative purposes [139]. For instance, reimbursement procedures defined by Medicare (the US social security organisation) requires the *coding* of a patient's stay into a series of Diagnosis Related Groups (DRG), which are derived from the International Classification of Diseases (ICD) [212] and Current Procedural Terminology (CPT) [25] codes. These codes are attributed to each patient after hospital discharge and therefore present several drawbacks:

1. each code refers to a specific hospital stay and cannot be linked to a specific ICU stay when multiple codes are present;
2. it is well known that the different codes attributed to a patient are usually post-processed to optimize the amount claimed from Medicare;
3. time stamps provided with procedure codes are only precise to the day, preventing any fine-grained analysis of their impact on patients' physiology.

Table 2.3 presents the three most common codes of primary diagnosis for the different ICU service types, illustrating the difference in patients, care and expertise that can be found in different ICU types.

Laboratory results are mapped onto a terminology called Logical Observation Identifiers Names and Codes (LOINC) that represent a universal standard for identifying medical laboratory results [193]. It contains more than 30,000 items that are specific to the site of collection of the blood sample, providing different codes for blood sugar measured

Table 2.3: *Description of the three most common main diagnoses at admission (primary ICD-9 codes) according to service type.*

Service	ICD-9	Description	Occurrence
CCU	410.71	Subendocardial infarction initial episode of care	625
	414.01	Coronary atherosclerosis of native coronary artery	348
	410.11	Acute myocardial infarction of other anterior wall	312
CSRU	414.01	Coronary atherosclerosis of native coronary artery	1838
	424.1	Aortic valve disorders	503
	410.71	Subendocardial infarction initial episode of care	379
MICU	038.9	Unspecified septicemia	607
	518.81	Acute respiratory failure	381
	507.0	Pneumonitis due to inhalation of food or vomitus	230
SICU	431	Intracerebral hemorrhage	440
	430	Subarachnoid hemorrhage	229
	434.91	Cerebral artery occlusion unspecified with cerebral infraction	142

for example, from an artery, the finger-tip or the ear-lobe. A novel global ontology, Systematized Nomenclature of Medicine (SNOMED) [67], is meant to replace these different components and seems to prevail in recent HIS designed by the industry. A fundamental change, next to the broader domain it covers, is that SNOMED is designed to be used throughout the patient's stay and feed the EMR, while ICDs were designed to extract information from the EMR after discharge to meet the secondary use of data purposes (epidemiology and reimbursement) [264].

2.2 Population Study

2.2.1 Identification of the cohort of interest

We have defined in section 1.7 the population of interest for this work as all adult patients with severe sepsis presenting hypotension for which adequate interventions were found (as per guidelines introduced in Section 1.5). Identification of the right popu-

lation is a key element of retrospective studies on clinical databases, and in particular when clinical definitions lack specificity to allow for computerization. Specifically, the definition of sepsis introduced in section 1.4 is full of vague vocabulary (*some*, *probable*, *suspected*, *presumed*, *looking*, *substantial*), which certainly cannot be translated into computer-readable criteria. Conversely, although SIRS constitutes a reproducible criterion, it is far too sensitive and lacks specificity. Last but not least, documenting infection from the lab results present in the database may be possible but would still discard all patients who *look* septic but do not have a positive culture.

Epidemiologists and health economists have long been exposed to this problem and have developed acceptable strategies to deal with it using some of the aforementioned administrative codes. Indeed specific ICD and DRG codes exist for sepsis but suffer from important drawbacks and have so far been dismissed by the scientific community [229, 84, 27]. Currently accepted strategies consist of looking at combinations of specific codes indicating infection or organ failure. In order to estimate the variation between these approaches, we extracted and compared the following cohorts:

- Cohort I: using single ICD-9 code for septic shock (785.52);
- Cohort II: combination of ICD-9 and CPT codes following according to Angus' definition [6];
- Cohort III: combination of ICD-9 and CPT codes following Martin's definition [189].

Figure 2.1 represents the size and overlap of populations identified with different criteria. Cohorts I, II and III respectively includes $1,034$, $6,970$, and $3,295$ patients accounting for 3.8%, 21.7% and 10.2% of all adult ICU admissions. These prevalences are consistent with literature that reports figures from 8.7% for septic shock [12] to 30.0% for sepsis and severe sepsis [289, 215]. Interestingly, 98% of patients in cohort I are also included in Martin's group (II), which in turn is included in cohort III with a similar overlap of 95%. These findings are in agreement with reports by Martin et al. [189].

The identification of sepsis population from single ICD-9 codes of septic shock was

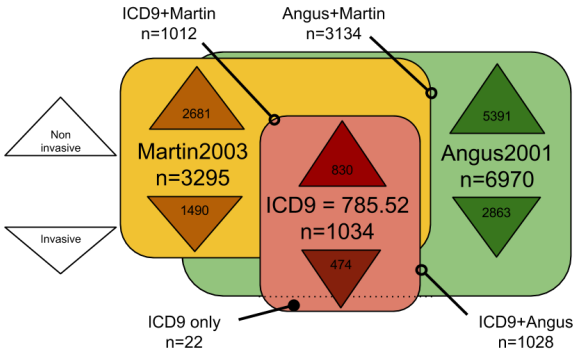


Figure 2.1: *Three different definitions of a population of interest extracted in the MIMIC-II database: ICD-9 codes of septic shock (785.52), Angus [6], and Martin criteria [189]. The plot shows the overlap between the different definitions and displays the number of patients at the intersection. The triangles indicates the counts for invasive (downwards triangle) and non-invasive (upwards) blood pressure measurements.*

also found to be biased towards a more severe population as demonstrated by Whittaker et al. [300], which also corroborates our findings. Angus' and Martin's criteria on the other hand target the sepsis population which does not necessarily have hypotension. In addition to this, administrative codes are related to hospital stay and it is unclear during which ICU stay the sepsis may have occurred when the patient has several admissions to the ICU. Thankfully, the identification of organ failures that are derived from interventions such as mechanical ventilation or Renal Replacement Therapy (RRT) can be linked to a specific ICU stay through the date used to code the intervention. Last but not least, using the chart events available in MIMIC-II, additional rules were created to identify precisely the time at which sepsis-induced hypotension occurred. The definition of hypotension was chosen to include patients who did not have invasive measurements of blood pressure and was defined as: at least 2 consecutive nurse-verified recordings of mean arterial blood pressure below 60 mmHg. This cut-off on ABP_{Mean} was chosen based on the study by Dünser et al. [80] that showed that the time spent below mean arterial pressure of 60 mmHg in the ICU was the best predictor of severity-adjusted mortality. Finally, the initiation or the increase of fluid and/or vasopressors up to three hours before the onset of hypotension was required to include the patient in the study. To conclude, the population in the study were all adult ICU patients with sepsis-induced hy-

potension and organ failure, for whom adequate treatment (fluids or vasopressors) was administered.

The use of SIRS criteria in addition to the administration of antibiotic therapy has been reported in clinical literature and constitutes an interesting alternative towards population selection. The SIRS criteria alone is over-sensitive, while more than 90% of MIMIC patients meet its definition at least once during their ICU stay. The use of an additional criteria such as the administration of antibiotic therapy, which can be identified from the Provider Order Entry (POE) table in MIMIC-II, could potentially narrow down to the population of interest. Using such criteria would essentially capture all patients for whom predefined antibiotics were ordered. More recently, the information related to blood culture was added in the database including sensitivity analysis of different antibiotics with respect to a detected micro-organism. Arguably, the use of antibiotic is an overly sensitive criteria (the drug is often given preventively) and the presence of a positive blood culture is under sensitive (most septic patients do not show a positive blood culture). Yet, antibiotic use and microbiology information all provide precise temporal information allowing for finer analysis of the sequence of events, which administrative codes do not offer. Consequently, they constitute interesting approaches towards population selection that could advantageously be explored, compared and contrasted in future studies.

The mortality rate of the population described in this work is 28.9%, which seems consistent with the selection of a population having a slightly lower severity than a population of severe sepsis patients. More precisely, the inclusion criteria did not require the presence of organ dysfunction (in addition to hypotension and use of vasopressor) and, shortly before the hypotensive episode, only 61% of patients in the cohort showed signs of organ failure. Patients who responded positively to the treatment for hypotension were purposely included in the cohort in order to potentially capture what, in their response to hypotension and its treatment, could have indicated a better outcome as opposed to patients who later evolved to have septic shock.

2.2.2 Description of the population studied

2.2.2.1 Population extracted from the clinical database

Figure 2.2 presents the flow chart for patient selection: 26% of all adult ICU patients did meet the Angus criteria for sepsis ($n = 6,970$) and $5,760$ had hypotension as defined above. About half of these patients received treatment for hypotension adding up to 8.2% of all ICU admissions. Finally, 14 patients were discarded from the analysis because they were missing more than 50% of the selected covariates (described in section 2.3) leaving $2,143$ patients for final analysis.

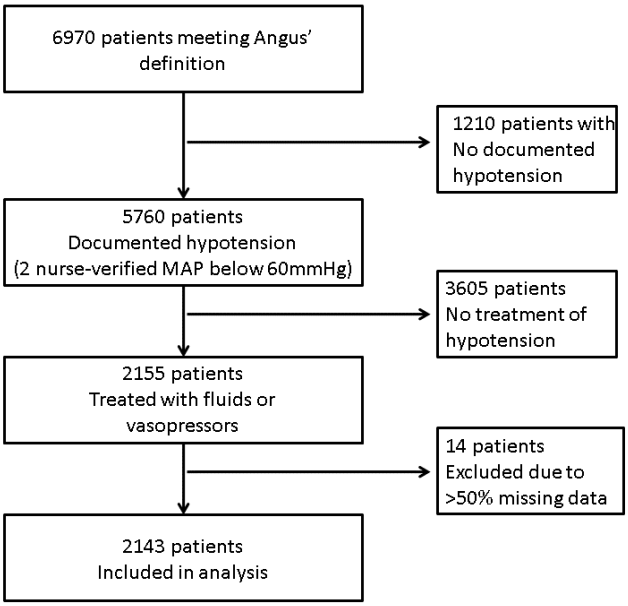


Figure 2.2: *Using the MIMIC-II database, we identified $2,155$ patients matching the ICD-9 and procedure codes defining severe sepsis according to Angus et al. [6], that met the definition of hypotension (at least 2 nurse-verified MAP recordings below 60 mmHg), and required fluid resuscitation or vasopressors. Fourteen patients with more than 50% missing data were excluded leaving $2,143$ for data analysis.*

Table 2.4 describes the surviving and non-surviving groups (pertaining to in-hospital mortality) at admission and over the hypotensive episodes with covariates showing a highly statistical significance difference ($p < 0.001$) in distribution between the two groups – using the Kolmogorov-Smirnov (KS) test. It shows that non-survivors are on average admitted with worse indicators of severity at admission : physiological (SAPS-I),

neurological with the Glasgow Coma Scale (GCS), cardiovascular (ABP_{Mean}), and overall amounts of organ failure. Interestingly, hypotensive episodes in this population occur much later (24.3 hours against 13.0 for survivors), are statistically significantly longer ($l = 2$ hours against 1.5 hour) and are associated with more aggressive treatments (Vasopressors, RRT, and mechanical ventilation).

Finally, to characterize the severity of hypotension in the extracted population, the volume and type of fluids administered to patients were extracted and are presented in table 2.5: (a) from admission to onset of the hypotensive episode, (b) one hour before the onset of hypotension and, (c) during the hypotensive episode. It shows that normal saline is by far the most common type of treatment used, while a third of patients still receive an equivalent amount of lactate ringers.

2.3 Description of the data extracted

ICU data is characterized by a very large number of covariates that include numerous laboratory results, vital signs, demographic and administrative information. The current paradigm in most medical studies is to select a subset of clinically meaningful covariates that are identified by one or more experts. This is particularly true for RCTs where the exact type and nature of data collected has to be detailed before asking for ethics committee approval. As a consequence, a number of confounding factors and unknown predictive variables may simply be missing in the analysis. On the contrary, the clinical data mining approach is to extract as many variables as possible and automatically identify those of interest. Ideally then, all item IDs present in all the tables should be extracted and used as single variables. It is common however that multiple item IDs are given to describe the same variable. In a context of unlimited data points, potentially necessary additional parameters may adequately be identified in order to account for the slight variation in meaning these additional item IDs might have. However, with the population described in the previous section ($n = 2,155$) it is preferable to merge as many

Table 2.4: *Characteristics of population studied (N=2,143) showing admission values and describing the hypotensive episode for surviving and non-surviving groups.*

Parameters	Survivors[a] (n=1,508)	Non-Survivors[a] (n=605)	P-values[b]
			At admission
Age (years)	68.5 (56.0-78.1)	70.6 (58.3-81.8)	<0.001
Gender (M, %)	52.7	54.7	0.39[c]
SAPS-I	16 (12-19)	19 (15-23)	<0.001
GCS	11 (6-15)	11.5 (7-15)	0.49
APS	33 (24-45)	47 (35-60)	<0.001
APACHE-III	47 (35 -60)	61 (48 -75)	<0.001
SOFA	1 (0-8)	8 (0-12)	<0.001
Van Walraven	9 (3-15)	12 (7-18)	<0.001
Elective Adm. (%)	13.2	4.0	<0.001[c]
Bypass Surgery (%)	14.2	3.5	<0.001[c]
Mean Arterial Blood Pressure (mmHg)	66.5 (62.5-72.0)	64.5 (61.0-70.0)	<0.001
Organ Failure (%)	34.8	49.9	<0.001[c]
			Hypotensive episode and treatments
Hypotensive episode onset (hrs)	13.0 (5.3-40.8)	24.3 (5.9-83.1)	<0.001
Length of hypotensive episode (hrs)	1.5 (1.0-3.0)	2.0 (1.0-3.8)	0.002
Crystalloid administered during hypotensive event (L)	1.8 (0.8-3.5)	2.0 (1.0-4.0)	0.028
Sedatives used (%)	59.7	71.8	<0.001[c]
Vasopressors used (%)	61.9	81.6	<0.001[c]
Renal Replacement Therapy (%)	12.5	25.7	<0.001[c]
Mechanical Ventilation (%)	36.1	52.2	<0.001[c]

[a] Continuous values presented as median ($25^{\text{th}} - 75^{\text{th}}$ quantiles);
[b] The p-value shows the result of a non-parametric Mann–Whitney U test with the null hypothesis that two groups have similar values against the hypothesis that one population has larger values;
[c] The p-value shows the result of a chi-square test of independence for binomial variables.

Table 2.5: *Different types of fluids administered to patients: average volume of fluid given over three time windows (from admission to onset, 1 hour before onset and during the hypotensive episode) and proportion of patients receiving it.*

	Colloids Mean Volume (L) (% of total patients)	Normal Saline Mean Volume (L) (% of total patients)	Lactate Ringer's Mean Volume (L) (% of total patients)
From admission (a)	0.782 (6.0%)	1.602 (56.4%)	2.063 (28.0%)
One hour before onset (b)	0.407 (1.9%)	0.834 (38.3%)	1.026 (16.8%)
During hypotensive episode (c)	0.518 (6.4%)	0.967 (71.0%)	1.065 (29.6%)

common item IDs as possible, sometimes across different tables. This requires clinical expertise for the data analysed. The data presented in this section is extracted for each patient and includes: outcome, demographics, Chronic Health Condition (CHC), physiological data (vital signs and laboratory results), waveforms, minute-by-minute trends (when available), and interventions.

2.3.1 Definition of outcome

The aim of this work is to improve the existing estimation of severity for severe sepsis patients. In order to do so, definition and extraction of the *outcome* is necessary. In general, an outcome can be a disease, sign or symptom like ARDS, hypotension or arrhythmia but is often chosen to be mortality or LOS for models estimating patient severity. Depending on the typical time-scale of the condition studied, different types of mortality can be considered to highlight better differences between surviving and non-surviving populations. Indeed, if five-year mortality is a relevant endpoint to explore the efficacy of a cancer treatment, it is rather inappropriate for studying risk factors during an acute condition like haemorrhage or shock. Outcomes selected for ICU-related conditions are generally ICU, in-hospital or thirty-day mortalities depending on the kind of prospect considered: economic, organisational, or medical. In-hospital mortality is a broader definition that includes patients for whom care has been withdrawn and who have been discharged to the ward with Comfort Measures Only (CMO) or Do Not Resus-

citate (DNR) codes. For instance, the difference between in-hospital and ICU mortalities varies from 2.5% in CCU to 5.8% in CSRU as shown by table 2.2 suggesting different approaches toward complications. In this work we consider in-hospital mortality as the outcome for the reason given earlier because this outcome is consistent with existing literature. We will call "positive" a patient who is not surviving hospital stay because we are looking at prediction of mortality. Likewise, a patient surviving will be counted as a negative case.

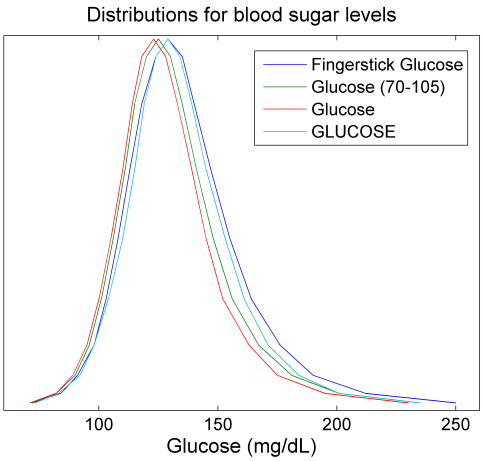
2.3.2 Physiological data

Physiological variables were historically the first variables considered for the estimation of patients' severity [15, 159, 173]. Yet, the lack of accuracy that these models developed over a decade, required not only re-estimation of the model's coefficients, but also the introduction of novel predictive variables such as CHC, demographics and admission data. The relative contribution of a variable or a group of variables, which we call the explanatory power, can be estimated with several techniques, introduced in section 3.1.1. The explanatory power of acute physiology at admission in Acute Physiology and Chronic Health Evaluation (APACHE) dropped from 73.1% in the 90's [163] to 65.6% in the late 2000's [313, 203]. A similar pattern has also been reported by Moreno et al. [203] by the SAPS team with more dramatic figures from 66% [174] to 27.5% (first hour after admission) [203] between 1993 to 2005, respectively. One plausible explanation for this drop is that it could reflect the improvement in care delivered to ICU patients over two decades: the proportion of mortality imputable to physiology effectively has dropped as a result of an improved management of acute conditions [201]. The differences observed in figures reported by different groups may be affected by geographical variations as well as the exact time of data collection (first hour for SAPS against first 24 hours from admission for APACHE). Despite this downward trend, physiological variables still constitute an essential source of information and should therefore be included in our study.

Figure 2.3: *(Left): Candidate variables for the extraction of glucose showing item IDs, label, occurrence in database, and sampling rate. (Right): Distribution of some of these item IDs for our population of patients.*

itemid	Label	Category	Number of ICU stay	Average sampling rate (per day)*
811	Glucose (70-105)	Chemistry	30,349	3.2
1529	Glucose	Chemistry	23,489	3.0
807	Fingerstick Glucose	Chemistry	20,937	5.0
3447	Glucose Monitor	None	5,875	1.1
3745	BloodGlucose	Quick Admit	1,442	1.1
3744	Blood Glucose	Chemistry	350	1.1
1310	Fingerstick glucose	None	3	1.7
1455	Fingerstick Glucose	None	3	1.3
2338	Finger stick glucose	None	2	1.0
2416	Finger stick glucos.	None	1	1.0
1812	ABG: glucose	None	1	1.0

* Days without a measurement were not considered for computing the mean.



From the technical standpoint, physiological data can be extracted from two main tables in MIMIC-II: laboratory and chart events. The former is connected to the laboratory information system that records all data related to blood or tissue samples, which are all coded with Logical Observation Identifiers Names and Codes (LOINC). The chart events table gathers all nurse-verified information including lab results, bedside monitor values (vital signs), and clinical observations (skin colour, mental and physical status). A fundamental difference between the two tables is that the first one uses an international ontology (LOINC) while the other relates to bespoke item IDs chosen by the nurse at the time of information key-in: for instance, searching for "glucose" in nurse labels leads to eleven distinct candidates as presented on the left hand side of Figure 2.3. The selection of correct labels to be merged requires a certain degree of expertise for the identification and selection of candidates, which can mostly be made from the sampling rate and prevalence as illustrated with the last two columns of the table seen on Figure 2.3. The description of all physiological variables extracted from MIMIC-II is given in table A.3, which presents the details of data extraction rules (table, foreign unique key, value and time attributes) as well as the selected item IDs.

Most physiological measurements will be taken at different times during the patient's stay. Some values like vital signs can be sampled as often as several times an hour, while

each sample comes with temporal information. More precisely, all rows in the chart and laboratory tables provide two time stamps:

Charttime corresponds to the time of observation;

Realtime corresponds to the time the information has been entered into the bedside monitor.

For instance a blood sample drawn at 11:45AM (Charttime) may not return from the laboratory before 2:30PM (Realtime). The former indicates the time at which a patient's physiology is observed while the latter relates to when blood levels are read and how they may influence the patient's care. Similarly, automated measurements are streamed from bedside monitors and verified by the nurse at different times. Unless stated differently, the chart time was always used since it better reflects the patient's physiology. The high temporal granularity available in MIMIC-II allows us to explore the evolution of these variables with respect to specific events, therefore the exact time-window at which physiological variables will be extracted will vary from study to study.

2.3.3 Neurological status

It is a generally accepted that a large part of relevant clinical observations are not collected in the HIS. In fact, how a patient *looks* often says more about his health than any number. For instance a patient reading a magazine in his bed with a systolic blood pressure below 60 mmHg certainly should be given a lower severity than another patient lying unconscious with the same pressure reading. In fact the prognostic value of neurological markers has long been noticed: for instance delirium is a strong predictor of mortality in mechanically ventilated patients [86, 85, 105]. In order to capture such information and reduce as much as possible the inherent variability of such observations different metrics were developed; two of them are available in the MIMIC-II database. The GCS is a clinical score ranking from 3 to 15 where higher values indicate a healthier neurological status. The GCS decomposes into: best verbal, motor, and eye opening

responses scoring between 1 and 4, 1 and 6, and 1 and 5 [268], respectively. Often however, the estimation of levels of consciousness is impaired by the use of sedation, which in turns requires pseudo-objective ways of assessing its efficacy. To do so different sedation-agitation scales were developed and evaluated. The Riker's Agitation-Sedation Scale (RASS) ranks from 1 (Unarousable) to 7 (Dangerous Agitation) with a neutral level at 4 (Calm and cooperative) [66, 241].

2.3.4 Microbiology results

There is long-standing evidence that the presence of specific micro-organisms, the site of infection, and the sensitivity of antibiotic therapy relates to severity of the infection [194, 6, 101, 60, 225, 256]. The latest version of the MIMIC-II database (v2.6) also includes results from microbiology and indicates the identified micro-organism, the site of sample collection, and the sensitivity to different antibiotics tested with the following label: resistant (R), intermediate (I), sensitive (S) and pending (P). Clearly not all combinations of organism, site, antibiotic, and sensitivity could be derived into single variables without increasing the number of covariates beyond reasonable proportions. Instead, we have identified the eight most commonly found organisms where resistant strains are as listed in Table 2.6. Sites and location were then individually coded as separate variables on four levels indicating the absence (0) and the presence (1) with sensitive (2), intermediate (3) and resistant (4) in-vitro sensitivity to antibiotic therapy.

2.3.5 Medication and intervention

Medications and interventions have never been used for the estimation of severity because they reflect local practice more than patients' physiology, which is generally avoided in benchmarking and evaluation of case-mix. Yet, the use of interventions can indicate chronic or acute dysfunction: presence of Renal Replacement Therapy (RRT) reflects a degree of kidney dysfunction that may not necessarily be captured by other physiological variables. In addition to this, while most severity scores focus on patients' physiol-

Table 2.6: *Eight most common organisms and locations showing resistive strains in the population.*

Mirco-organism Description	Patients (%)	Location Description	Patients (%)
Staphylococcus aureus	24.5	Sputum	22.6
Enterococcal species	13.2	Blood culture	15.9
Staphylococcus coagulase Neg.	9.6	Urine	14.5
Escherichia coli	6.9	Swab	14.0
Pseudomonas aeruginosa	6.7	Catheter IV	7.6
Klebsiella pneumoniae	4.1	MRSA Screen	2.9
Enterococcus faecium	2.7	Bronchoalveolar	2.7
Enterobacter cloacae	1.8	Tissue	2.2

ogy at admission, we believe that response to treatments and interventions may bring additional discriminative power. As a result, extraction related to treatments and interventions with the highest temporal resolution possible was sought. Management of sepsis-induced hypotension is characterised by strict guidelines that offer a homogeneous framework for the study of patients' response. Treatments typically can be split into two categories: those with relatively short action-time whose effects can be observed within a few minutes or up to several hours (fluids and vasopressors) and others that typically spread over days (glucose management, mechanical ventilation, and RRT).

Table A.1 describes the different labels identified from the database for the extraction of treatments: different types of fluids, vasopressors, sedatives, and insulin. For fluids the procedure was complicated by the presence of a large variety of labels and associated ITEMIDs. The procedure to identify the correct ones was:

1. a preliminary list of names was collected from four clinicians (Paris, New York City, and Boston);
2. searching terms ("D5", "NS", "saline", ...) were identified and applied to all labels;
3. the resulting list was then reviewed by all clinicians iteratively until agreement was reached;

4. finally, ITEMIDs were split into different categories: normal saline, colloids, lactate ringers, dextrose in saline, and hypotonic solutions (rare).

Similarly, RRT was extracted from the procedure table with the description field matching "dialysis" (ITEMID $\in \{3895, 3927, 3942, 3943, 3995, 5496\}$). Finally, the presence of mechanical ventilation was identified with the presence in the chart events of ITEMIDs specific to mechanical ventilation settings: ventilator mode, number and type with ITEMIDs 720, 721, and 722, respectively. While this approach is recommended, it ignores all patients arriving in the ICU with mechanical ventilation but without later change to the settings. To account for these, mechanical ventilation was also extracted from the procedure codes matching "mechanical ventilation" (ITEMID $\in \{9390, 9670, 9671, 9672\}$) and discharge summaries matching regular expressions: $CPAP$, $mech.\{0, 10\}vent$, and $intubat$. A regular expression is a sequence of symbols that codes for specific patterns of characters: for instance the regular expression $r = .a\{1, 3\}b$ matches any sub-string of length between 2 and 5 characters, starting with a and finishing with b, like in "abrupt", "arabesque", "rabbit" or "grab". The three extraction methods were stored as independent variables to compare the predictive ability of each.

2.3.6 Demographic data

Demographic data comprises all high-level information about a patient that is considered as constant for a hospital admission: age, gender, ethnicity. Additional administrative codes may also improve estimation of severity such as admission type (elective, emergency, or urgent) and source (referral, Emergency Department (ED), transfer from hospital or other health institution).

2.3.7 Chronic health conditions

CHC conditions are now an essential component of modern severity scores in the ICU and are now used as covariates in most of them [158, 313, 131, 203]. Demographic data

and CHC constitute baseline risk factors that are independent of the quality of care patients receive in the ICU: nothing can be done about a patient's age, ethnicity or type of admission. Recent severity scores, solely based on CHC [57, 83, 295] (see section 3.2.1 for a complete description) clearly demonstrate the importance of these variables in terms of explanatory power for in-hospital mortality. Earlier reports however suggest that the six CHC included the APACHE-III score only account for 2% of the total explanatory power [158] when used in addition to physiological components rising to 5% in the later version of the score [313]. The same trend, although with a larger percentage, which was reported in Moreno et al. [201] shows that the relative importance of CHC discriminatory power has risen from 4% to 49.9% for the SAPS score [174, 203]. Studies of the APACHE cohorts suggest that CHCs may have a stronger contribution on the population of patients with sepsis [162].

Interestingly, not all CHCs are associated with a higher risk of mortality: obesity and drug abuse are two examples of CHCs usually associated with a better outcome. The protective effect of being slightly overweight was recently demonstrated by Flegal et al. [91] in a meta-analysis ($n = 2.88$ millions) relating standardized Body Mass Index (BMI) groups to all-cause mortality. Other studies also reported on the protective effect of being overweight in the ICU for all kinds of mortality considered [1]. Yet there is little medical literature to explain this phenomenon and it was recently suggested that this could be an artefact called the *obesity paradox* that challenges the validity of BMI as a valid measurement for obesity [283]. Moreover, in the ICU a non-obese patient with a heart rate of 90 beat per minute (bpm) and respiratory rate of 20 Breaths per minute (bpm) certainly is sicker than a high-BMI patient with the same heart rate and respiratory rate for whom baseline physiology is different [148]. Finally, Elixhauser et al. [83] suggest that seriously ill patients could not be given codes for non-threatening conditions simply because of work overload; as a consequence, their presence could be a surrogate for a relatively healthier patient with a lower risk of in-hospital death.

Beyond such considerations, these covariates are required to replicate state-of-the-

art models for prediction of mortality with which our work will be compared. Accurate identification of comorbid conditions from the MIMIC-II database is not straightforward, firstly because ICD codes lack precision, but also because coding guidelines do not recommend the coding of chronic health if no impact on care and procedures is expected [212]. The information about chronic conditions was therefore extracted from both administrative data and free-text discharge summaries and then stored as distinct variables.

2.3.7.1 From administrative data

Extraction of comorbidities from administrative data was done according to Elixhauser et al. [83] who describe a series a ICD-9 codes matching 31 comorbidities. In addition to this, hepatic failure, Acute Immune Dysfunction Syndrome (AIDS), cirrhosis, myeloma and metastatic cancer were extracted from unique ICD-9 codes that are given in table A.4.

2.3.7.2 From discharge summaries

The approach chosen for the identification of comorbidities in discharge summaries was to look at some specific regular expressions. It is not uncommon however that a mention of a CHC is made without particular reference to the patient: for instance the chronic conditions of close relatives are mentioned when clinically relevant ("father has cirrhosis") and these constitute an important source of false positives. Similarly, CHC can be specifically ruled out or simply hypothesised ("suspicion of …", "…denies presence of …"). In order to account for these, the twenty closest words to each positive match were scanned for negative terms and the case was dismissed if relevant. The left-hand side of table 2.7 presents the positive and negative regular expressions that were identified for each CHC. Positive terms were determined by experts (drug names) and negative ones were selected after careful examination of a representative list of positive cases among which false positives were found. Because this procedure is iterative, and does not implement any kind of cross-validation, we do not claim it to be generalizable. Moreover,

Table 2.7: *Description of regular expressions used to scan discharge summaries in order to find positive cases of CHC that are required to compute APACHE-IV [313]. Negative terms refer to regular expressions used to search for false positives within the twenty closest words to the positive match. The right hand side of the table (4^{th} column onwards) presents the performance of the automated extraction compared to expert labels.*

CHC	Positive terms	Negative terms	FP	FN	TP	TN	Sen	Spe
Lymphoma	'lymphom'	'no', 'not', 'father', 'mother'	11	0	1	438	1.00	0.98
Leukemia	'leuk'	'no', 'not', 'father', 'mother'	0	0	0	450	N.A.	1.00
Metastatic cancer	'metastasis', 'metast.{0,10}cancer'	'no', 'not', 'father', 'mother'	23	1	3	423	0.75	0.95
AIDS	'HIV','AIDS'	'no', 'not', 'father', 'mother', 'NSAIDS', 'negative', 'test', 'hearing', 'shivering', 'denies'	7	0	2	441	1.00	0.98
Hepatic Failure	'hep.{0,10}fail', 'liver.{0,3}fail', 'lactulose'	'no', 'not', 'father', 'mother'	73	0	2	375	1.00	0.84
Cirrhosis	'cirrhosis'	'no', 'not', 'father', 'mother'	49	1	7	393	0.88	0.89
Immunosuppression	'immu.{0,10}sup','tacrolimus', 'prograf', 'cyclosporine', 'neoral', 'sandimmune', 'mycophenolate', 'mofetil', 'cellcept', 'sirolimus', 'belatacept', 'nulojix', 'atg', 'anti-thymocyteglobulin' ,'basiliximab' ,'simulect' ,'methotrexate' ,'leflunomide' ,'etanercept' ,'infliximab' ,'adalimumab' ,'golimumab' ,'certolizumab', 'anakinraand', 'tocilizumab', 'abatacept', 'rituximab', 'chemotherapy'	'no', 'not', 'father', 'mother'	43	1	8	398	0.89	0.90

Abbreviations: AIDS: Acquired ImmunoDeficiency Syndrome, FP: False Positive, FN: False Negative, TP: True Positive, TN: True Negative, Sen: sensitivity, Spe: specificity.

terms found in discharge summaries are likely to reflect some local practice and would not necessarily be replicated in a different type of hospital or in a different region.

In order to validate this approach in our data, five hundred discharge summaries were randomly selected and manually reviewed by an ICU consultant working at BIDMC. To facilitate the task, a GUI (illustrated in Figure 2.4) was developed to present experts with quotes from the discharge summary matching the regular expressions aforementioned. The tool was developed in Matlab and made publicly available online so that it could easily be modified to deal with various sources of free text and include new types of

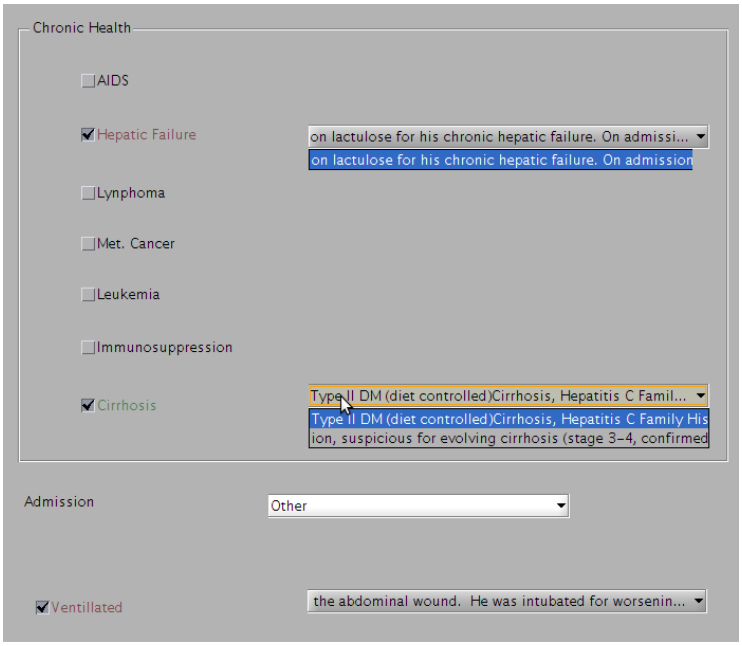


Figure 2.4: *Graphical User Interface (GUI) developed for the manual review of discharge summaries. On the left-hand side of the GUI, the complete discharge summary is provided (not displayed on this figure). The area of the GUI visible in this figure represents the right-half of the GUI, in which the occurrences of identified chronic health items are printed for the reviewer for confirmation. A colour code (red/green) indicates agreement with data extracted from administrative tables: Cirrhosis for instance shows green while "Hepatic failure" is printed in red.*

information to extract. At the end of this process both False Positives (FPs) and False Negatives (FNs) were presented a second time to reviewers offering them an opportunity to revise their decision. The right-hand side of table 2.7 presents the performance of the automated approach with expert labelling used as the reference. Finally, the tool also implemented collaborative features allowing different experts to share the work or process the same discharge summaries.

2.4 Pre-processing

Data pre-processing is an integral part of the data modelling and evaluation procedures that will be detailed in sections 3.1.2 and 4.1. It is generally considered good practice to avoid any pre-processing derived from metrics extracted from the data used to estimate final performance. Physiological variables however present properties that are equally valid in different databases allowing for universal pre-processing steps. Such

Table 2.8: *Identification of low and high cut-off values for the rejection of physiologically impossible observations computed from standard deviation, 0.01^{th} and 99.99^{th} percentiles compiled from all ICU stays.*

Description (unit)	0.1^{th}	99.9^{th}	Std	Low cut-off	High cut-off
Invasive ABP_{Mean} (mmHg)	49	131	18.1	30.9	149.1
Non-Invasive ABP_{Mean} (mmHg)	46	119.7	15.8	30.2	135.5
Heart Rate (bpm)	53	180	36.1	16.9	216.1
Respiration Rate (bpm)	9	38	67	2.3	44.7
Temperature (^oC)	34.9	38.9	1.5	33.4	40.4

pre-processing was implemented as described by Johnson et al. [145]: identification of specific artefacts (HR$= 0$, weight $= -1$), conversion of measurements entered with wrong units (^oF instead of ^oC), and rejection of physiologically impossible values. For each variable x, the range of physiologically possible values was defined between cut-off$_{low}(x)$ and cut-off$_{high}(x)$ defined as

$$\text{cut-off}_{low}(x) = p_{0.01\%}(x) - \sigma(x) \tag{2.1}$$

$$\text{cut-off}_{high}(x) = p_{99.99\%}(x) + \sigma(x) \tag{2.2}$$

where $p_{0.01\%}(x)$ and $p_{99.99\%}(x)$ denote the 0.01 and 99.99% percentile of all values from all $40,426$ ICU admissions, respectively and $\sigma(x)$ the variable standard deviation for for the same population. To illustrate, table 2.8 shows the cut-off values for the main physiological variables. The inspection of other variables also indicates that this unusual combination of standard deviation and percentiles offers a reasonably reliable criteria for the automated identification of non-physiological values.

Chapter 3

Baseline for prediction of mortality

We have briefly introduced the purpose and nature of severity scores for the critically ill (section 1.7). Then we presented in section 2.3 the data available and extracted from the MIMIC-II database in the light of covariates required to compute these scores. In this chapter we will set a baseline to which our results will be compared. In the first place, we will briefly describe the technical background that is common to most of these techniques. Then, general severity scores designed on hospital and ICU populations will be further explained, implemented, applied to our data and customised. Other severity scores that are specific to the severe sepsis population will also be introduced. It will be shown however that comparing or replicating these studies cannot easily be achieved; the performance of these approaches as reported in literature will be presented. The performance of various scores on the data used in this thesis will serve as a baseline for direct comparison of future results. The performance reported in the literature from other databases will serve as an indirect point of comparison.

3.1 Design, evaluation & comparison of severity scores

To date, a large majority of the articles dealing with the identification of risk factors (i.e. the estimation of patient risk of mortality) have used similar techniques, namely logistic regression with minor variations in its implementation. Likewise, the evaluation

of these models consistently relies on a few metrics including the area under the receiver operating curve and the Hosmer-Lemeshow goodness-of-fit criteria. The details of these techniques will be introduced in this section to help the understanding of previous work.

3.1.1 Introducing simple modelling techniques

3.1.1.1 Logistic regression model

Logistic regression relates a dichotomous dependent variable, denoted y, to K independent predictors x_j where $j = 1 \dots K$. In this model, the conditional probability of the outcome for observation i given the data – for instance the probability of patient i not surviving his hospital stay given his admission variables – is denoted $Pr(y_i = 1|x_i) = \pi(x_i)$ where π is the logistic function expressed as :

$$\pi(x) = \frac{1}{1 + e^{-g(x)}} \tag{3.1}$$

$$g(x) = \beta_0 + \sum_{j=1}^{K} \beta_j x_j \tag{3.2}$$

Conversely, the probability of patient i surviving is expressed by $Pr(y_i = 0|x_i) = 1 - \pi(x_i)$. Assuming all observations (patients) are independent, it is possible to define a function describing how likely are a set of parameters β_j given the data collected. This proportion is named the likelihood of the parameters and is defined as follows:

$$\mathcal{L}(\beta|x) = \prod_{i=1}^{N} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{(1-y_i)} \tag{3.3}$$

Maximization of the likelihood is a key concept of logistic regression analysis. For mathematical convenience, the log of equation 3.3 is usually considered for parameter estimation leading to the following definition of the *log-likelihood*:

$$\log \mathcal{L}(\beta|x) = \sum_{i=1}^{N} y_i \pi(x_i) + (1 - y_i)\,(1 - \pi\,(x_i)) \tag{3.4}$$

From equation 3.4 can be derived K partial derivatives with respect to the coefficients , which cancel where the log-likelihood maximizes. These equations, called the *likelihood equations*, can be solved computationally with the help of iterative techniques for instance. The estimated coefficients using the Maximum Likelihood Estimation (MLE) technique are usually noted with a hat symbol, which usually denotes the most likely outcome:

$$\hat{\beta} = \underset{\beta}{\text{argmax}} \ \log \mathcal{L}(\beta|x) \tag{3.5}$$

3.1.1.2 Interpretation of logistic regression models: variable importance

Parameter estimation using the MLE technique also provides information about the standard error of the identified coefficients $\hat{SE}(\hat{\beta})$. The latter can be derived from the *observed information matrix* (see Hosmer Jr et al. [135, p37-38]) and used to estimate how significant a model coefficient is. More precisely, the Wald-test relates a maximum-likelihood estimate to a z-statistic assumed to be normally distributed:

$$z = \frac{\hat{\beta}_j}{\hat{SE}(\hat{\beta}_j)} \tag{3.6}$$

P-values reported from the Wald-test are some of the most reported statistics for studies where a dichotomous outcome is related to covariates by the mean of a logistic regression model. Yet, the interpretation of these p-values is subject to caution and clinical expertise should always prevail. Indeed, it is not uncommon to violate assumptions during the process of model design. For instance, the MLE of logistic regression coefficients generally assumes normality, non multi-colinearity of data, as well as independence between observations. Clearly clinical data violates these simple assumptions: physiological variables are not necessarily normally distributed, some are certainly correlated, and patients admitted at the same time in the same service are certainly not independent since they share the same resources. Clearly these elements should raise awareness on the necessity to subject any result to clinical expertise.

Another interesting aspect of the logistic regression model is the possibility to relate a change in a covariate to an increase in risk estimate. This feature certainly contributes to the popularity of logistic regression models over more complex non-linear techniques. As detailed by Hosmer Jr et al. [135, p50-51], the increase in risk estimate associated with an increase of c units for the i^{th} covariate is called the Odds Ratio (OR) and can be described as:

$$\hat{OR} = e^{c\beta_i}. \tag{3.7}$$

For dichotomous variables ($c = 1$), the odds ratio can be seen as the increase in risk associated with the presence of the covariate. For instance, in the context of in-hospital mortality prediction, an OR of 1.1 for variable "metastatic cancer" indicates that a patient with this comorbidity is 10% less likely to survive his hospital stay. The benefit in clinical understanding of the model when variables are not pre-transformed is obvious. For instance, an increase in heart rate of 10 beats per minute may translate to a relative increase in risk of mortality of $\frac{p(dying|HR+10bpm)}{p(dying|HR)} = 1.2$. When variables are pre-transformed however such presentation of the coefficients does not necessary provide a better clinical understanding of their meaning.

Another approach towards the estimation of the relative importance of a covariate is presented by Knaus et al. [158]: the explained variance (χ^2, see next section) of the full model is compared to that without the covariate, as well as the explained variance of a model only including the covariate. Ratios of these variances indicate the actual and maximum relative contributions of each covariate.

3.1.2 Evaluation and comparison of model performance

An essential part of model validation is the estimation of the "performance"; or, how well does the model achieves what it was designed for? Doing so necessarily requires a more precise definition of "performance", which certainly is context-dependent. Essentially, metrics of performance tend to reflect three properties : significance, calibration, and

discrimination.

3.1.2.1 Statistical significance of logistic regression

Because logistic regression is solved through MLE, a natural metric to estimate how well the model fits the data (i.e. the *fit*) is the probability of the parameters given the data provided or the likelihood $\mathcal{L}(\beta) = P(\beta|x, y)$. Unfortunately, the likelihood is a data-dependent metric and more standardized metrics are desired.

Deviance is a simple test that compares a model against the hypothesis that its coefficients are zero. To do so, it takes the ratio of the log-likelihoods of the full model and the model with all coefficients set to zero, which defines the deviance:

$$D = -2 \log \left[\frac{\mathcal{L}(\beta_i = 0)}{\mathcal{L}(\beta_i)} \right] \tag{3.8}$$

The resulting D-statistic follows a $\chi^2(n-k-1)$ distribution that translates to a measure of statistical significance. This metric measures the overall significance of a model whilst it does not assess the individual contribution of each coefficient: to do so techniques introduced earlier such as the Wald-test will be preferred. More generally, likelihood ratios of nested models (of increasing complexity) can be used to compare different models and help choose the right balance between model performance and complexity.

3.1.2.2 Goodness-of-fit

Goodness-of-fit is an indication of how well estimated probabilities of the outcome are related to the observed outcome. Risk stratification is a typical example of a clinical application that requires a good calibration: a subgroup of patients predicted with higher risk must show a higher incidence of the outcome.

Pearson χ^2 statistic If a model generates a continuous output for a dichotomous outcome (such as mortality), it is possible to define clusters of patients (whose severity score

falls between two given values) and then to compare *observed* against *estimated* risk within these groups. The χ^2 statistics is then defined as follows:

$$X^2 = \sum_{g=1}^{10} \frac{(o_l - e_l)^2}{e_l} \tag{3.9}$$

$$o_l = \sum_{l \in D_l} y_i$$

$$e_l = \sum_{l \in D_l} P(x_i)$$

where D_l, $l = 1 \dots g$, denotes the observations in the l^{th} decile of risk, and o_l and e_l represent the observed and estimated risks within these deciles, respectively. The distribution of the χ^2 statistic is compared to a chi-square with $g - (k-1)$ degrees of freedom (valid for $p+1 > g$ where p is the model dimension and g the number of bins used to compute the statistic). However the actual number of degrees of freedom has not been assessed formally [176] and situations such as when $g \approx k$ lead to instability of the test [135, p. 157].

Hosmer-Lemeshow goodness-of-fit test Lemeshow and Hosmer [176] present a review of statistics for estimating the fit of logistic regression and introduce a new statistic that accounts for the limitations of the χ^2 statistics. The Hosmer-Lemeshow $\hat{C}^*$-statistic ($\text{HL}_{\hat{C}^*}$) is similar to the χ^2-test [313, 176] but splits the estimation of risk according to the actual class of the observations. The score, which will further be referred to as the $\text{HL}_{\hat{C}^*}$, was defined as :

$$\hat{C}_g^* = \sum_{k=0}^{1} \sum_{l=1}^{g} \frac{(o_{kl} - e_{kl})^2}{e_{kl}} \tag{3.10}$$

where,

$$o_{1l} = \sum_{i \in D_l} \hat{y}_i \qquad o_{0l} = \sum_{i \in D_l} (1 - \hat{y}_i)$$

$$e_{1l} = \sum_{i \in D_l} \hat{y}_i \qquad o_{0l} = \sum_{i \in D_l} (1 - \hat{y}_i)$$

and o_{1l} and e_{1l} are the observed and expected risks of positive cases falling in the l^{th} decile of risk, respectively; likewise o_{0l} and e_{0l} represent the observed and expected risks for the negative cases, respectively. The $\hat{C}^*$ statistic is assumed to follow a χ^2 distribution with degrees of freedom $g - 2$ if $p + 1 < g$. According to Hosmer Jr et al. [135] deciles of risks of equal size guarantee a better fit to the χ^2 distribution than fixed thresholds and will therefore be preferred. The number of groups g, is usually set to $g = 10$, leading to 8 degrees of freedom, but should be updated when the number of variables in the model is too important (if the hypothesis $p + 1 < g$ does not hold). Despite its wide use in estimation of model calibration, the $\text{HL}_{\hat{C}^*}$ presents an obvious limitation. For instance when p is great and n small, the high number of groups and small number of patients will lead to very small (or null) values of e_{kl} and therefore infinite or very high values of the statistic. Similarly, Kramer and Zimmerman [167] have demonstrated that badly calibrated models tend to provide a statistic that is linearly related to the sample number. Furthermore recent discussion by Hosmer Jr et al. [135, p. 160] reports doubts on the validity of the number of degrees of freedom used for the test and confirms that the reported statistic follows a $\chi^2(g - 2)$ distribution. Interestingly, when the fit is assessed on external data, authors have used an additional 2 degrees of freedom [135, p. 204-5] to account for the "loss of one degree of freedom in each of the probability intervals" [269, p. 210].

The standardized mortality ratio or SMR is reported for scores predicting mortality. It is simply defined as the ratio between the predicted mortality rate over the observed mortality rate. An SMR lower than 1 indicates that the model underestimates severity

whilst a ratio greater than 1 indicates an overestimation of severity.

3.1.2.3 Discriminative power

Another aspect of model performance is the ability to discriminate between positive and negative outcome. Because the logistic regression model for estimation of severity provides a probability of dying, a threshold has to be selected above which observations are classified as non-surviving. This threshold is named the *operating point* and it transforms a continuous prediction into a binary output hence allowing derivation of additional metrics of performance. The prediction has then to be compared with the true outcome defined in section 2.3.1.

Discrete measures of discriminatory power At every operating point (a threshold on the predicted probability) it is possible to allocate each observation to a predicted group (positive or negative). According to the combination of observed and predicted outcomes, each observation is defined as a True Positive (TP), True Negative (TN), FP, or FN as detailed in Table 3.1. With these definitions, a patient correctly classified as non-surviving is a true positive, while a patient correctly identified as surviving is a true negative. From these numbers can be derived additional metrics of discrimination :

- Positive Predictive Value (PPV), the proportion of correctly identified cases within all cases predicted positive;
- Negative Predictive Value (NPV), the proportion of correctly identified cases within all cases predicted negative;
- Sensitivity (Sen), the proportion of correctly identified cases within all observed positive cases;
- Specificity (Spe), the proportion of correctly identified cases within all observed negative cases;
- Accuracy (Acc), the proportion of correctly identified cases within all cases.

Table 3.1: *Relation between observed and predicted outcome showing the confusion matrix in the middle square that defines: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Additional measures of discriminatory power are given on the right and bottom of the table as combinations of these.*

		Observed outcome $\oplus$	Observed outcome $\ominus$	
Predicted Outcome	$\oplus$	TP	FP	Positive Predictive Value $PPV = \frac{TP}{TP+FP}$
	$\ominus$	FN	TN	Negative Predictive Value $NPV = \frac{TN}{FN+TN}$
		Sensitivity $Sen = \frac{TP}{TP+FN}$	Specificity $Spe = \frac{TN}{FP+TN}$	Accuracy $Acc = \frac{TP+TN}{TP+TN+FP+FN}$

Area under the receiver operatoring curve (AUROC) When assessing the overall discriminatory power of a technique, it is sometimes interesting not to restrict the analysis to a specific operating point but rather to consider a continuum of thresholds. While doing so, sensitivity and specificity can be defined as a function of the operating point (op) and it is possible to plot $Sen(op) = 1 - Spe(op)$ for all possible values of op. Such a plot is called the Receiver Operating Characteristic (ROC) and indicates the possible combinations of $(Sen, 1 - Spe)$. The line defined by $Sen(op) = Spe(op) = 0.5$ is the diagonal joining the bottom left hand-side corner of this plot to the top right-hand side corner, which represents the discriminative power of chance (any random guess). Integrating the ROC gives the area under the ROC curve, or Area Under the Curve (AUROC), which indicates how much discriminatory power a model can provide over all possible thresholds. Interestingly, the AUROC has been shown to have similarities with the Wilcoxon statistic W [118]. Consequently, the AUROC can be interpreted as the probability of any positive case ($x_i, y_i = 1$) being predicted with a higher probability $\hat{y}_i$ than any negative case ($x_j, y_j = 0$) so that $AUC = W = Pr(x_i > x_j)$. This probabilistic view of the AUROC provides a more comprehensible understanding of its real meaning and offers alternative ways to compute its value, which is particularly helpful for low sample size.

3.1.3 Comparison of model performance

The development of new severity scores requires the estimation, not only of the performance, but also of how a model can compare to another; and how statistically significant is the difference. The ratio of log-likelihood introduced earlier can effectively compare nested models of increasing complexity (meaning covariates included in the simplest model are also present in the more complex one) to provide a measure of statistical significance. In terms of discriminatory power, the comparison of two AUROCs computed from two different models on the same cases (paired statistics) will be detailed in this section. Finally, more recent attempts for model comparison such as the Net Reclassification Improvement (NRI) and the Integrated Discriminative Improvement (IDI) will be introduced.

3.1.3.1 Comparing discriminatory power

Evaluating the difference between two AUROCs and how significant is the difference is not a straightforward issue. In order to do so, Hanley and McNeil [118] have demonstrated how the Standard Error (SE) of an AUROC can be estimated as follows:

$$
\begin{aligned}
SE &= \sqrt{\frac{W(1-W) + (n_{\ominus}-1)(Q_1 - W^2) + (n_{\oplus}-1)(Q_2 - W^2)}{n_{\ominus} \times n_{\oplus}}} \\
Q_1 &= \frac{W}{2-W} \\
Q_2 &= \frac{2W^2}{1+W}
\end{aligned}
\tag{3.11}
$$

where $n_{\ominus}$ and $n_{\oplus}$ denote the number of patients in the negative and positive groups respectively;Q_1 is the probability of any two randomly chosen positive observations being predicted with a greater probability than any observation and Q_2 denotes the probability of one randomly chosen positive observation being predicted with greater probability than two randomly chosen negative cases [43, 118]. Subsequently, the SE of the

difference between two AUROCs is defined by Hanley et al. [119] as:

$$SE(W) = \frac{W_1 - W_2}{\sqrt{SE_1^2 + SE_2^2 - 2 \cdot R \cdot SE_1 \cdot SE_2}} \tag{3.12}$$

and assumed to follow a standard normal distribution from which statistical significance can be obtained. This implementation introduces the correlation coefficient R between two models that are derived from the same cases. This coefficient is estimated from a lookup table provided by Hanley et al. [119]. Another implementation of this problem suggested by DeLong et al. [77] offers a solution that does not require the use of a lookup table.

3.1.3.2 Net reclassification index

The NRI first introduced by Pencina et al. [221] has gained interest in the medical community as a complementary metric to asses the performance of a new model with respect to an older one. It first introduces the notion of upward movement (*up*) as a change from a lower category of risk with the old model to a higher category with the new model and the downward movement (*down*) as its opposite. When dealing with models predicting mortality, a patient previously classified as non-surviving and predicted to survive in the new model would be considered as a *downward* move. The NRI is therefore defined by equation 3.13

$$\begin{aligned} NRI = &(P(up|y=1) - P(down|y=1)) \\ &- (P(up|y=0) - P(down|y=0)) \end{aligned} \tag{3.13}$$

$$z = \frac{NRI}{\sqrt{\frac{P(up|y=1)+P(down|y=1)}{n^+} + \frac{P(up|y=0)+P(down|y=0)}{n^-}}} \tag{3.14}$$

where n^+ and n^- denote the number of positive and negative outcome, respectively. Pencina et al. [221] show that a simple asymptotic test for the null hypothesis of $NRI = 0$ can be used (equation 3.14).

3.1.3.3 Integrated discriminative improvement

The NRI provides a metric of performance at a specific operating point and an equivalent metric is suggested by Pencina et al. [221], which accounts for all possible classification thresholds. The IDI that is defined by equation 3.15 where IS denotes the integration of sensitivity over all the cut-off values and IP is the equivalent for "one minus specificity". As for the NRI, a simple asymptotic test can be used to test the null hypothesis H_0 : $IDI = 0$ (equation 3.16).

$$IDI = (IS_{new} - IS_{old}) - (IP_{new} - IP_{old}) \tag{3.15}$$

$$z = \frac{IDI}{\sqrt{SE^2_{events} + SE^2_{nonevents}}} \tag{3.16}$$

3.1.4 Choice of metric performance

All the metrics introduced above reflect a combination of discriminatory power and calibration, essentially reflecting how well the model fits the data. Ultimately, the choice of metric depends on the expected use for the model. For instance, if a model predicting mortality was accurate enough to forecast a single patient's outcome, measures of discriminatory power would certainly be the primary metric to compare these models. The AUROC offers an aggregated measure of discrimination that is informative for the data scientist, although real applications require the identification of a specific operating point. Depending on the type of intervention and the associated risk of predicting false positive or false negative, different metric like sensitivity, specificity, PPV or NPV can be considered. For example, a model intended to drive care withdrawal would preferably be evaluated with PPV: care must not be discontinued for patients who actually have a chance to survive. Conversely, the decision to use a specific low-cost treatment showing few side effects would target models with high sensitivity because the benefit of treating any positive case would potentially outweigh the risk of giving the drug to many negative patients. Because the use of models predicting mortality is so far limited to risk strat-

ification and benchmarking, optimal calibration is preferred to discrimination. In fact, a performance metric can (and must) be as specific as possible. For instance, a model designed to identify the 10% at-risk patients for whom increased level of care could be provided should be evaluated with a measure that solely considers the highest decile of risk rather than an integrated calibration metric over all deciles. Finally, it is worth noting that most of these metrics will be correlated since a model offering excellent calibration is unlikely to show poor discrimination.

In order to estimate the behaviour of these different metrics, an artificial data set equivalent in size to that introduced in section 2.2.1 ($N = 1000$) was generated. The dataset was generated using the logistic function as described in equation 3.17 where x was randomly generated from a normal distribution with zero mean and unit variance. The two parameters $c = 0 \dots 5$ and $\sigma = 0 \dots 2$ were used to artificially generate lack of calibration and fit (or a combination of both). The perfect model was defined by $y_{true} = y(0, 0)$ with associated targets $t_{true} = y_{true} > 0.5$ that were used as a baseline to compare models with non-null c and σ. The predicted probabilities $y(c, \sigma)$ were used in combination with t_{true} to compute performance metrics and y_{true} to compute comparison metrics with associated p-values.

$$y(c, \sigma) = \frac{1}{1 + exp^{-(c+x+N(0,\sigma^2))}} \tag{3.17}$$

3.1.4.1 Results and discussion

The results of this study are presented in figure 3.1 showing the evolution of the metrics against levels of noise σ and calibration c on the x and the y-axis, respectively. As expected, AUROC values show a strong horizontal gradient and are insensitive to lack of calibration. Interestingly, the Hosmer-Lemeshow $HL_{\hat{C}^*}$ and the Log-Likelihood ($\log \mathcal{L}$) show equivalent behaviour and capture both lack of calibration and discrimination, which is why such metrics should be preferred over metrics of discriminatory power, such as the AUROC. The $HL_{\hat{C}^*}$ offers the advantage of providing a p-value but our simulations

showed that it could not identify adequately the model of perfect fit and calibration, instead preferring one with a low level of noise. Additionally, visual examination of figure 3.1 indicates that the log-likelihood is more stable and should be a preferred metric for comparison or optimization. In order to quantify this, we define a "roughness" metric, $\mathcal{R}$, indicating how much local gradient is present for slight changes in model fit. $\mathcal{R}$ was estimated by calculating the median of the discrete Laplacian (sum of second derivatives) computed at all values of (c,σ) computed with the following $2D$ filter:

$$D^2_{c\sigma} = \begin{bmatrix} 0.5 & 1 & 0.5 \\ 1 & -6 & 1 \\ 0.5 & 1 & 0.5 \end{bmatrix} \tag{3.18}$$

$$\text{and the roughness is given by } \mathcal{R} = \text{median}_{c,\sigma}\left(D^2_{c\sigma}\right) \tag{3.19}$$

This criterion confirms and quantifies the initial visual feel and supports the use of $\log \mathcal{L}$ over the $\text{HL}_{\hat{C}^*}$ as an evaluation and optimization criterion with roughness levels of $\mathcal{R}_{LL} = 0.05$ and $\mathcal{R}_{HL} = 0.15$. Finally, in terms of comparative metrics, the IDI will be preferred for similar reasons: it captures changes in discriminatory power as well as calibration, while offering the most stable metric gradient ($\mathcal{R}_{IDI} = 0.04$).

Two techniques were introduced in section 3.1.3.1 for the estimation of the statistical significance of the difference in discriminatory power between two models. P-values given by each technique are noted p_h and p_d for Hanley et al. [119] and DeLong et al. [77], respectively. Their comparison reveals nearly perfectly overlapping distributions and extremely good correlation (Pearson correlation $r = 0.86$) where the Pearson correlation coefficient is defined by

$$r(W^h, W^d) = \frac{\sum_{i=1}^{N}(W^h - \bar{W^h})(W^d_i - \bar{W^d})}{\sqrt{\sum_{i=1}^{N}(W^h - \bar{W^h})^2}\sqrt{\sum_{i=1}^{N}(W^d_i - \bar{W^d})^2}} \tag{3.20}$$

and $\bar{W}$ denotes the mean of W. Consequently the Delong method was used in this thesis from hereon, since it allows for a full analytical solution.

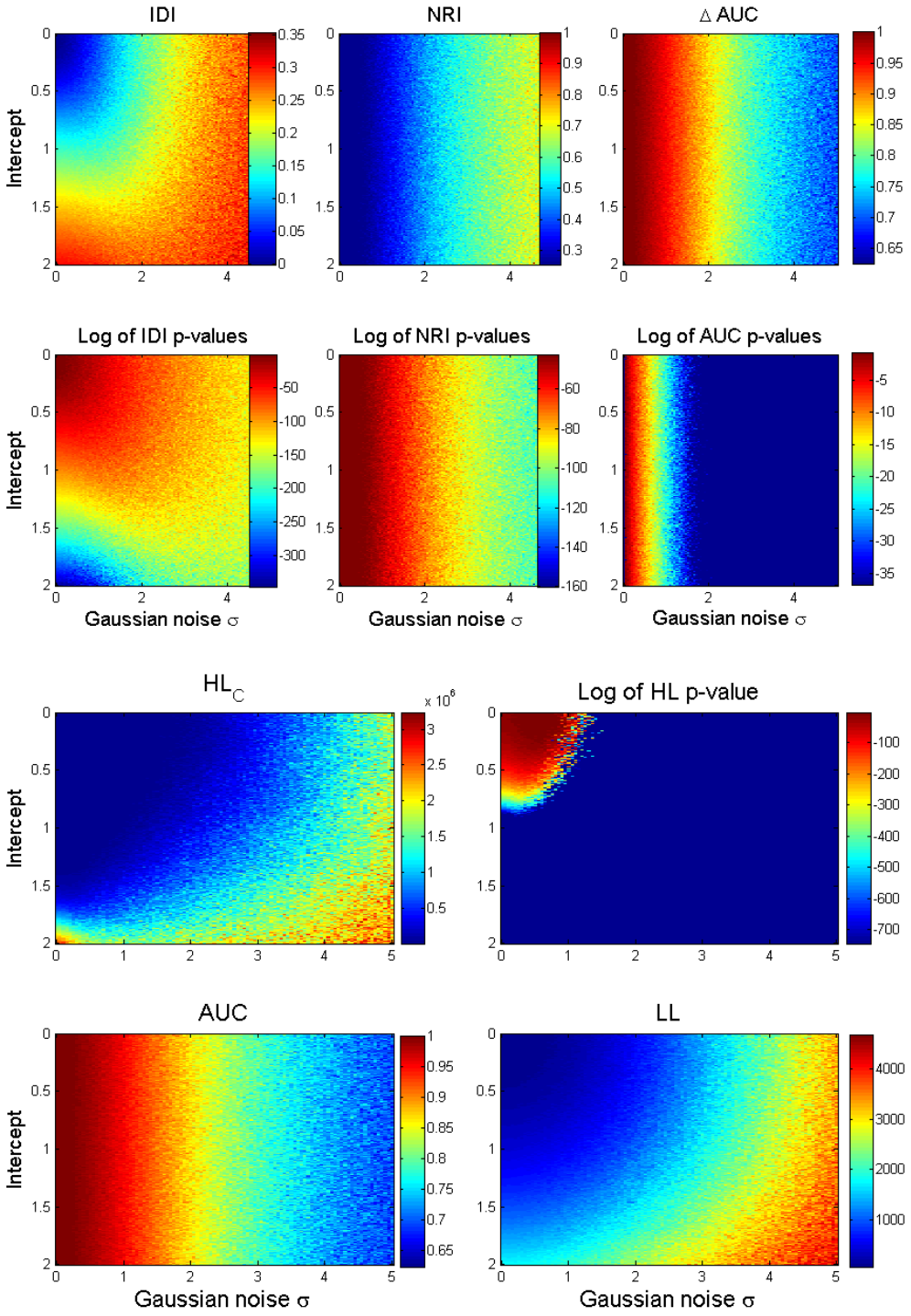


Figure 3.1: *(TOP): Three metrics of model comparison with different levels of lack of calibration (vertical axis) and noise (horizontal axis): NRI, IDI and AUROC. The two models being compared are equivalent in the top left-hand region and diverge in terms of both calibration and discrimination as the noise power and intercept increase (towards the lower right-hand side). (BOTTOM): Evolution of metrics of performance Log-Likelihood (*log $\mathcal{L}$*)*

3.1.4.2 Estimating performance on out-of-sample data

The gold standard for model validation is to collect new data prospectively and to compute the aforementioned metrics of performance on predictions made with this data. Unfortunately in practice prospective data collection is not always possible. Pseudo-prospective approaches consist of mimicking this behaviour in a retrospective setting: the most recent observations are kept aside during design for model validation. However, the exact sequence of patient admission is lost during the process of de-identification of the MIMIC-II database, which makes this approach impossible. In this context, *cross-validation* is the best option for model validation [172]. To do so, the data set is usually split into independent and complementary datasets:

training-validation set, or *design* set, noted X_{design} and composed of training $X_{training}$ and validation $X_{validation}$ data;

test set, noted X_{test}.

Note that the proportion ν between the design and test sample size may greatly influence the result. Furthermore, performance will vary depending on what training set is drawn from the data, as well as to which validation set it is applied. For instance missing values in the training data may impair parameter estimation. Similarly, different design/test distributions may affect results. As a consequence it is desired to quantify this variability and more elaborated cross-validation techniques involving statistical *re-sampling* offer this option:

K-Fold cross-validation defines K independent and complementary groups of equal size. The model fitting is repeated K times, each time leaving one k^{th} of the data aside for model validation. At the end of the procedure each observation has been used equally for training (K times) and test (once). This guarantees equal contribution of all observations leading to the computation of K performance metrics.

Bootstrapping draws with replacement B random training sets consisting in ν% of the entire data [172]. An estimate of the metric of performance and its variability can

thereby be computed, albeit with two limitations: the results are not reproducible and the respective contribution of each observation is uncontrolled.

Jack-Knifing is similar to bootstrapping with the difference that, for a given ν, all possible combinations of data split are considered. This offers a reproducible statistic at the expense of computational cost. In this sense, bootstrapping is seen as an approximation to the Jack-knifing estimate.

Leave-One Out is a special case of Jack-Knifing where ν is chosen so that a single observation is left out for validation. N training datasets of size $N-1$ are used for model design leading to N independent predictions that are used for model estimation. Unfortunately, this technique does not allow the estimation of the variance of the result but is still sometimes preferred in data-limited contexts.

3.2 Existing strategies for the estimation of severity

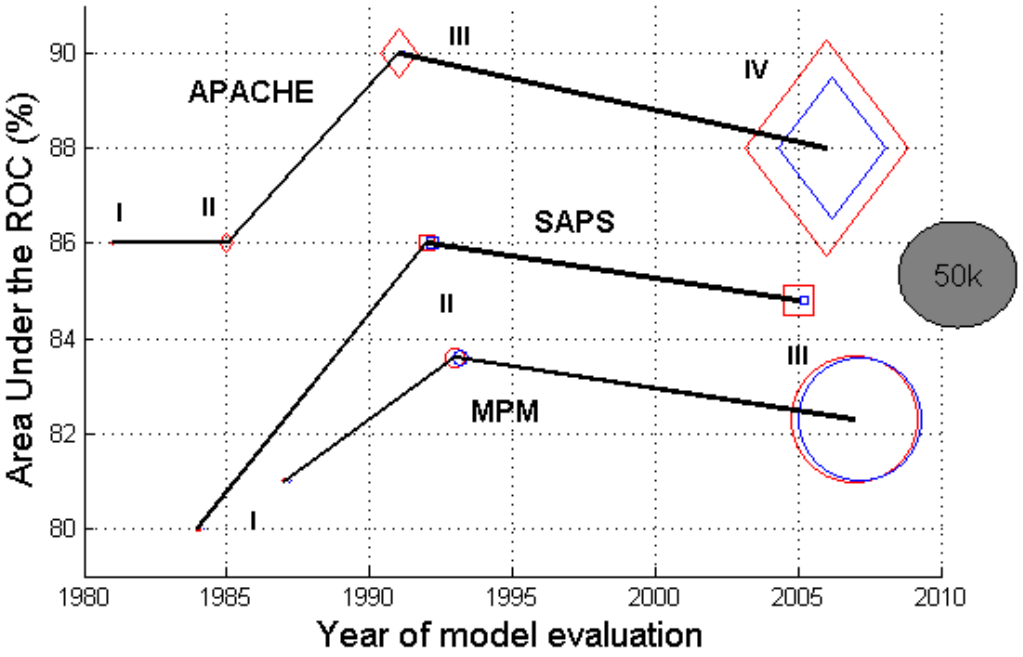
The technical background introduced in the previous section covers most of what has so far been attempted for prediction of mortality in general populations of patients. The main models we introduce in this section are APACHE, SAPS, and Mortality Prediction Model (MPM). Figure 3.2 illustrates the time at which different versions of these scores were developed with respect to their performance (AUROC), in relation to sample size (design and validation when available). Other relevant models of severity will be presented. Finally, some of these will be implemented and adapted to our population.

3.2.1 Previous attempts at predicting severity

3.2.1.1 Acute physiology and chronic health evaluation

The first version of the APACHE score was presented by Knaus et al. [159] in 1981 and clearly targets risk stratification applications over individual predictions, mostly for evaluation of new treatments in populations of critically-ill patients. The score is composed of

Figure 3.2: *Evolution of model performance (AUROC) over different versions of the scores for APACHE (diamonds), SAPS (squares), and MPM (rounds). For each model and each version, the log of the relative size of the training (red) and test (blue, if it exists) sample size is indicated. The grey circle indicates half a million patients for comparison.*



an admission and a physiologic component. The Acute Physiology Score (APS) includes 34 variables collected during the first 32 hours following admission to the ICU and is meant to capture different physiological systems: cardiovascular, respiratory, renal, gastrointestinal, haematological, neurological in addition to presence of sepsis. Cut-off values and points were defined by a panel of experts with varying degrees of expertise. The chronic health score defines four categories ranking from good prior health (A) to severe restriction of activity due to disease (D). The final estimation integrated the two components in a multiple logistic regression analysis together with additional covariates such as age, sex, and pre-admission status. The final prediction of mortality was evaluated on 582 university hospital admissions leading to a Pearson chi-squared statistics of $\chi^2 = 11.18, p < 0.02$ (see equation 3.9). The first version of the APACHE score was an elaborated proof-of-concept severity score, which - according to a later report from the authors [160] - was too complicated to be externally validated and consequently support clinical research on a large scale.

An updated version of the score was therefore proposed by [160] including only 12 physiological variables that were collected during the first ICU day. Age, type of admission (diagnosis, surgery/medical, and elective/emergency) and severe chronic health conditions (organ dysfunction or immunosuppression) were also included in the model. All covariates were used in a logistic regression model predicting hospital mortality. Coronary Artery Bypass Surgery (CABG) patients ($N = 785$) were removed because they represent a severe group (APACHE-II $= 12.4$) with low death rate (1.5%). This design resulted in a simple model for prediction of mortality covering chronic health, admission status and acute physiology. It was evaluated in 5,815 ICU admissions but results are however difficult to compare since only specificity, sensitivity, PPV, NPV, and accuracy were reported at different operating points. Integration of the ROC presented [160, Figure 5] indicated an AUROC of 86.7%. Even though it is unclear which cross-validation technique was implemented in this model, later external validations indicate that levels of overfitting were kept within a reasonable range [150, 249].

This version was again updated by Knaus et al. [158] to account for shifts in clinical practice, to benefit from larger sample size, and to add novel covariates such as pre-ICU LOS and more accurate admitting diagnosis. The final score was finally derived from $17,440$ patients from 40 North American hospitals. Novel features for this update included: the use of 112 admission diagnosis categories, a detailed acid-base disturbance score – pH-Partial Arterial pressure of CO_2 ($PaCO_2$) – and a combined neurological score (combination of GCS components). Again, all components were combined with a multiple logistic regression model with reported AUROC of 90%. No external validation has ever reported such levels of performance even though the model was consistently proven to perform better than its predecessors [310, 152]. A possible source of overfitting is the determination of some comorbidity coefficients from the entire database [158, p. 1621]:

> Because of the large number of disease categories in the APACHE III data base, coefficients for each disease classification with a sufficient number of patients and/ or deaths were obtained using the entire data file.

Surprisingly, this version of the score remained unchanged for fifteen years despite important changes in the practice of intensive care [38, 26, 128]. Eventually, version IV of the score was recently presented by Zimmerman et al. [313] as a mere update of version III although substantial improvements were proposed: larger dataset, additional covariates, and interaction coefficients. Indeed, APS coefficients were derived from 78,970 patients recorded between 2002 and 2003 for model design. Additional covariates including admission source, type, diagnosis (116 categories), and pre-ICU LOS were refined. Finally, use of restricted cubic regression splines was introduced in the score to account for non-linear interactions [120, p. 18-31]. The model was finally trained on 131,618 consecutive ICU admissions and evaluated on the following 52,647 patients leading to an AUROC of 88% and a good calibration ($\text{HL}_{\hat{C}^*} = 16.9$, $p = 0.08$). This model constitutes the state-of-the-art modelling technique for prediction of mortality in the ICU.

3.2.1.2 A simplified acute physiology score for ICU patients

SAPS-I was developed by [97] in 1984 and validated with data collected from 679 ICU admissions in France for whom vital status at hospital discharge was recorded. Fourteen physiological variables were manually identified by experts and combined with a score allocating points for each variable based on their value. For instance, a heart rate between 140 and 179 beats per minute corresponds to 3 points. The score is computed from data collected during the 24 hours following admission and taken as the sum of points derived from each variate. This model was reported to perform as well as the APS model [159] for a lighter data collection cost. Because variables, ranges, and scores are selected arbitrarily, it is possible that they accidentally optimized results on the available data, inducing slight overfitting. Fortunately, later validation of the score on external data [52] reported equivalent performance.

In the meantime, other models of severity introduced the use of statistical techniques for the choice and weighting of variables [158, 269, 178, 177] and an updated version of the score was introduced by Le Gall et al. [174]. This version was developed from $13,152$

patients that were randomly allocated to a design set ($8,369$) and a test set ($4,628 \approx 35\%$). The model used an equivalent scoring technique but this time the parameters were determined by an objective methodology instead of expert knowledge. The score was also converted it into a probability of hospital mortality with the use of logistic regression. Of 37 collected variables, only 17 were kept in the final model including physiological values, neurological status (GCS), chronic diseases, and type of admission. The final model was reported with an $\text{HL}_{\hat{C}^*} = 3.70$, $p = 0.86$ and an AUROC of 86% on the test set.

Finally, the later version of the score (SAPS-III, described by Moreno et al. [203]) was designed to account for various sources of heterogeneity present in the datasets: mostly local variations in practice. The data was collected from $16,784$ patients from 303 ICUs to capture: (1) patient characteristics before admission, (2) reason for admission, and (3) degree of physiologic derangement. The different thresholds for continuous variables were selected following a univariate analysis. The model was designed within a 5-fold cross-validation procedure during which variable selection was carried out following specific rules described by Moreno et al. [203]. The final model consisted of 20 variables that yielded to an AUROC of 84.8% and calibration of $H_{\hat{G}} = 10.56$, $p = 0.39$ (The $H_{\hat{G}}$ Hosmer-Lemeshow goodness-of-fit is a slight variation of the H_C test defined in Equation 3.10 whereby the deciles are of equal size rather then defined by equally distant cut-off values).

3.2.1.3 Mortality Probability Model

The MPM was initially introduced by Lemeshow et al. [177], a world-leading expert of logistic regression models [176, 135] who had previously been involved in the development of the SAPS model [174]. The rationale behind the first version of the model was to derive the model parameters from maximum likelihood rather than expert opinion. First, covariates were objectively selected, and then multiple logistic regression was used to provide a probability of hospital mortality. Two models were considered: MPM_0 computed up to one hour after admission to the ICU and MPM_{24} computed at the end of day

1. Each model included seven covariates for which coefficients were derived from 743 and 463 patients, respectively. Interestingly, the validation was carried out prospectively on 2,028 patients included after model design. Performance reported good calibration for MPM_0 ($HL_{\hat{C}^*} = 6.53$, $p = 0.7$) but not for MPM_{24} [269].

To correct the lack of fit of MPM_{24} as well as to account for "changes in ICU technologies, practice and populations" two new models were later developed on data collected from 12,610 patients that were split into two equal sets for design and validation. The MPM_0-II model is composed of 15 variables reflecting physiology, chronic and acute derangement at admission. The MPM_{24}-II model on the other hand is composed of 13 variables (5 at admission and 8 during the first 24 hours). Performance reported on the test set ($n = 6,514$) showed good calibration ($p > 0.05$) and discrimination (AUC>80%) for both models.

Subsequent external validations of MPM-II reported a degrading calibration and an overestimation of the risk of death. This phenomenon presumably originates in the increased ICU capacity for correcting acute physiological insults. The updated score MPM-III [130, 131] was designed from 50,307 ICU stays collected in 135 units from 98 hospitals (50% of the total dataset). New covariates were also introduced to represent groups of patients with especially low/high risk of death such as patients with gastrointestinal bleeding and full codes – including CMO. Non-linear interaction with age was tailored to reflect diminished physiologic reserve. The final model offered good calibration ($HL_{\hat{C}^*} = 11.62, p = 0.31$) and discrimination ($AUC = 82.3\%$).

3.2.1.4 Oxford Acute Severity of Illness Score (OASIS)

Recent efforts from the APACHE group have opened the way for machine-learning approaches towards model development. Previous severity scores were developed based on a combination of expert input and objective statistical rules. The most recent models have gained in complexity and have been evaluated with increasingly sophisticated validation procedures but still partly rely on expert input. Latest work from Johnson et al.

[146] includes fully-automated feature and weight selection. Using data from 72,474 patients collected from 49 US hospitals from 2007 and 2011 and machine-learning algorithms, a minimal set of physiological variables has been identified to predict severity of illness. In the end, the model is similar to previous severity scores - with ranges and associated points for each covariate - but it is simpler whilst including only ten covariates. The model was validated in a pseudo-prospective manner using 23,618 patients admitted between 2010 and 2011 showing good calibration ($HL_{\hat{C}^*} = 43.8$, $p = 0.08$) and discriminative power equivalent to previous models of greater complexity ($AUC = 83.7\%$ against $AUC = 82.2\%$ for APS).

3.2.1.5 Other severity scores

In addition to the mainstream severity scores presented above, other attempts have been made to provide an estimate of severity. These models are somewhat different to the former in terms of population size and complexity of techniques implemented. Yet some of them have been shown to achieve high performance and were found clinically to be useful.

Complete septic shock score A single attempt has been made by Baumgartner et al. [24] to design a prediction method specific to a population of patients with septic shock. The model was developed to predict mortality after the first day in the ICU with a multiple logistic regression model derived from $n = 45$ patients. The scoring system was developed from univariate analysis of variables selected by experts. Inclusion of variables in the final model was done after comparing variable distributions between survivors and non-survivors in two groups: the design group and the validation group ($n = 43$). In the end, two versions of the score were developed: the Simplified Septic Shock Score (SSSS) including only physiological variables and the Complete Septic Shock Score (CSSS) additionally including chronic health and type of infection. Three aspects of model design however require careful interpretation: small sample size, relatively large number of co-

variates and the use of the validation set during the design procedure. The combination of these three factors is known to lead to overfitting.

Chronic health conditions Recently, Elixhauser et al. [83] have introduced a score using binary inputs for 30 comorbidities that can be extracted automatically from administrative datasets using the ICD codes (version 9). Some studies combined these binary comorbidities using logistic regression models derived from the training data and generated a one-dimensional comorbidity score [96, 184]. The addition of 30 covariates during model design is however not necessarily desired, particularly in the presence of a limited dataset. Walraven et al. [295] suggested coefficients for each chronic health condition that were derived from $228,000$ consecutive hospital admissions. Interestingly, drug abuse and obesity were, again, found to decrease hospital mortality probability. Conversely, metastatic cancer, liver disease and lymphoma were the conditions associated with the most severe outcome. Another severity index was proposed by Charlson et al. [56] with expert defined weights for a reduced number of comorbidities. This index has been externally validated [55, 57].

Sequential organ failure assessment (SOFA) This score was introduced by Vincent et al. [284] in order to describe the degree of organ failure present in a patient. In this sense it is more a descriptive metric rather than a severity score. It was not developed to predict risk of hospital death and its use is consequently not restricted to the first 24 hours following admission unlike most scores described in this chapter. The scoring system was designed with expert knowledge and evaluated on 1,700 patients.

3.2.2 Other biomarkers of severity

Severity scores are by far the most widely available techniques for the estimation of risk of mortality. A large body of clinical publications however investigates the relation between other types of biomarkers and mortality. These biomarkers can be the concentration of

specific proteins, which are assumed to vary significantly during the course of infection and inflammation. Because such biomarkers are specific to the pathophysiology of sepsis, they are expected to offer a deeper representation of the patient's state. We present in Table 3.2 a selected subsets of such attempts.

3.3 Performance on population of sepsis patients

This section will provide a baseline with which to compare our results. This baseline is essentially composed of two parts: (a) the performance reported in the clinical literature that deals with different approaches towards prediction of mortality and (b) the direct application of available models on the population described in section 2.2.1. The latter offers a precise comparison point with same-case derived metrics. The former on the other hand elaborates on techniques that cannot necessarily be replicated retrospectively on the data used in this work (because data was not collected prospectively).

3.3.1 Performance reported in literature

Scores of severity have proven particularly useful in the field of sepsis for both research and clinical practice, including support for clinical trials [298, 37, 147] and risk stratification [254, 49, 147]. Attempts have been made to design models that are specific to a population of patients with sepsis syndrome. A comprehensive overview of such models is provided by Barriere and Lowry [22] who published work up to 1995. Later studies also fall into one of the three categories they identified: categorical risk factors (admission and procedures), endotoxin and cytokine serum concentrations, and sepsis scoring systems. Relevant studies are summarized in Table 3.2 that presents sample size, design, statistical techniques, performance, and most predictive variables for each study.

The best multivariate performance reported is found in [41] ($AUC = 86.9\%$) with a SOFA score calculated for $n = 329$ patients with shock (approximately a third of which was sepsis-induced). Likewise, [11] reports an AUROC of 86% using univariate changes

Table 3.2: *This table summarizes previous attempts at predicting mortality in patients with sepsis and severe sepsis (SSep) showing the variables or models of interest, the performance (reported statistic), the population size, and the mortality rate. In the Population Type column is indicated in parentheses whether the metrics reported are from a prospective (p) or retrospective (r) study.*

Variables (Model)	Statistic	Population Type	Population Size (Tr/Te)	Mortality Rate	Year - Reference
MODS	$AUC = 85.2\%$	SepS (p)	N.A./329	55%	2002 [41]
SOFA	$AUC = 86.9\%$				
McCabe score, SOFA, (LR)	$AUC = 86.3\%$	SepS (p)	189/N.A	58%	2000 [11]
lactates, cortisol	$HL_{Pval} = 0.44$				
Simplified Septic (LR)	$Spe = 82.0\%$	SepS (p)	45/43	51%	1992 [24]
Shock Score	$Sen = 84.0\%$				
MAP area $<= 65$ mmHg (LR)	$AUC = 85.3\%$	SepS (r)	111/N.A.	$30\%^a$	2005 [282]
SAPS-II	$AUC = 79.7\%$, $HL_C = 23.60$	SSep (p)	N.A./250	46%	2003 [16]
APACHE-II	$AUC = 78.2\%$, $HL_C = 34.89$			$60\%^a$	
MPM-II$_{24}$	$AUC = 82.3\%$, $HL_C = 29.79$				
Age, Lactates, (LR)	$AUC = 73.0\%$	SSep (r)	899/473	40%	2008 [240][d]
GCS	$HL_C = 0.69$				
ΔCortisol (CPHR)	$AUC = 61.0\%$	SepS (p)	218/N.A.	28%	2007 [73]
ΔCortisol/Albumin				$41\%^a$	
MOF	$p = 0.008$	SepS (p)	N.A./71	52%	1991 [18]
APACHE-II	$p = 0.013$				
MPM	$p = 0.091$				
APACHE-II (LR)	$p < 0.001$	SSep[b] (p)	112/N.A.	44.5%	2010 [19]
Albumin	$p = 0.009$				
Antithrombin-III	$Sen = 96.0\%$, $Spe = 76.0\%$	SepS (p)	60/N.A.	$32 - 77\%$	1992 [92]
protein-C	$Sen = 96.0\%$, $Spe = 76.0\%$				
MEDS	$AUC = 74.0\%$	SSep (p)	N.A./216	32.9%	2010 [70]
mREMS	$AUC = 62.0\%$				
CURB-65	$AUC = 59.0\%$				
RIFLE	$AUC = 67.8\%$, $\chi^2 = 5.2$	Sepsis (p)	N.A./121	47.1%	2009 [59]
Organ Failure Nbr.	$AUC = 74.2\%$, $\chi^2 = 2.0$				
APACHE-IV	$AUC = 71.0\%$, $\chi^2 = 5.3$				
SOFA	$AUC = 66.9\%$, $\chi^2 = 4.7$				
CURB65	$AUC = 78.0\%$	SSep (p[c])	N.A./419	22.6%	2007 [20]
CRB65	$AUC = 73.0\%$				
SIRS	$AUC = 68.0\%$				

[a], ICU mortality was reported instead of in-hospital mortality;
[b], Community acquired bloodstream infection;
[c], pneumonia, retrospective use of data collected prospectively;
[d], unpublished work by Richard and Lacson [240]

in cortisol levels after a corticotropin stimulation in $n = 189$ septic shock patients. Unfortunately, the procedure is not part of routine clinical measurements. Other biomarkers considered include: clot-waveform analysis [273], procalcitonin in post-operative patients with severe sepsis [72], and cortisol in relation to albumin [73].

Many studies also present univariate statistics and highlight the importance of a small number of commonly measured physiological variables: respiration, transaminases, bilirubin, and albumin; others such as Antithrombin-III and Protein-C [92] are less common. Clearly, univariate results should be handled with care - if considered at all - since they do not account for possible confounding variables. Finally, general severity scores presented in the previous section have also been externally validated on different populations of patients with sepsis, severe sepsis, septic shock, or a combination of these. Results presented cover the second versions of APACHE, SAPS and MPM [16, 18, 19, 273, 237] and SOFA [41, 11]. They demonstrate equivalent performance of these scoring systems and stress the need for customisation and recalibration.

Interestingly, some models focused on the estimation of severity not only at admission but also during the evolution of sepsis. For instance, shortly after the onset of SIRS, Hebert et al. [127] suggest a score reflecting multi-system organ failure (MSOF). It is composed of points for seven physiological systems that are used in addition to age for the prediction of 30-day mortality in 154 patients. The performance was however not judged to be good enough by the authors to use the prediction at an individual level. Similarly, Pittet et al. [225] presented a study investigating biomarkers of severity throughout the evolution of sepsis: at admission and during the onset of sepsis and early signs of organ dysfunction. In this case, the change in numbers of organ dysfunction was found to relate better to severity than the actual number of organ dysfunction. Finally, Bossink et al. [40] introduced a model to estimate the severity of 300 febrile patients and showed that a small number of covariates (temperature, GCS, respiratory rate, and albumin) were more strongly associated with mortality than SIRS during the first three days following admission to the ICU. Similarly the Pitt Bacteremia Score (PBS), first described by Chow

and Yu [60, Table 7], estimates severity at the time of the first positive blood culture and was successfully externally validated [237].

One high-quality study was performed by Knaus et al. [162] (of the APACHE group) in order to offer a reliable index of severity for patients with sepsis, which could be used to compare the efficacy of drugs. The model was designed with 28-day survival analysis from $1,195$ patients that were split into training, validation and test sets for feature selection, model design, and external validation. Interestingly, acute physiology still accounted for the largest part of model-explained variance (71%), while other less important factors contributed equally: chronic conditions and pH (6.3% each), age (5.7%), pre-ICU LOS (5.2%), and WBC (4.6%). The use of a Cox-proportional hazard ratio model was discouraged in this study [22] because the hazard function is assumed to be constant over time, which is violated by patients with sepsis [162]. The proposed estimate of risk of death successfully outperformed APACHE-II ($AUC = 76$% against 71%). A similar attempt was more recently presented by the SAPS group [204]. The model was developed according to latest definitions and views on sepsis, namely the concept of Predisposing conditions, the nature and extent of the Insult, the nature and amplitude of the host Response and the degree of concomitant Organ dysfunction (PIRO). Based on $2,628$ patients, a model was developed to reflect: Predisposition (age, admission variables, chronic health), Injury (acquisition, site, and organism), and Response (organ failure and dysfunction). A five-fold validation procedure led to good calibration ($\text{HL}_{\hat{C}^*} = 5.33$, $p = 0.87$) and discrimination ($AUC = 77.2$%) while SAPS-III showed discrimination of only 73% on the same cases. The size and quality of these studies builds a strong rationale for the development of a sepsis specific severity score.

From all the articles reviewed, Richard and Lacson [240] probably present the work that is the most similar to the approach we propose. Sixty-three variables were extracted during the first three days from admission from $n = 1,372$ severe sepsis or septic shock patients selected according to ICD-9 codes of severe sepsis in the MIMIC-II database. Features were selected using forward selection prior to the use of a number of modelling

techniques including logistic regression. The best AUROC (73%) on the test set ($n = 473$) was obtained with Logistic Regression (LR) using the following variables during the first 24 hours of stay: Age, minimum GCS, mean lactic acid, minimum O_2 levels, maximum potassium levels and maximum temperature. This work, unpublished to date, suffers from at least two important drawbacks. First, the use of single ICD-9 codes for sepsis and severe sepsis has long been dismissed by the scientific community (see section 2.2.1). Then, the performance is reported on a single test set, while good practice recommends more complex cross-validation procedures (see section 3.1.2).

Finally a striking observation whilst looking at table 3.2 is the remarkable heterogeneity of the studies presented in terms of:

Setting with retrospective, prospective, observational, and interventional studies;

Inclusion criteria different degrees of sepsis severity are considered and identification of these in the population is made according to different definitions;

Population different types of time periods, geographical areas, hospital, and ICU types for which baseline risks are highly variable;

Methods a broad range of covariates (biomarkers) are reported and model performance are assessed with metrics that is not always directly comparable.

In particular the method of extraction of groups of patients (study population) with severe sepsis or septic shock in retrospective studies varies greatly. Varpula et al. [282] used the equivalent of DRG codes for sepsis (sepsis, pneumonia, meningitis, or peritonis) in addition to the presence of vasopressors during the first 24 hours from admission leading to an incidence of 7.8% for a population of septic shock. Richard and Lacson [240] on the other hand used single ICD-9 codes for severe sepsis or septic shock from the MIMIC-II database leading to an incidence of 5% (smaller incidence for a less severe population). A more complex approach was chosen by Knaus et al. [162] who details a precise list of inclusion/exclusion criteria based on patients' lab results and procedure

records, leading to an incidence of sepsis of approximately 2%. This confirms our findings presented in section 2.2.1 that relate small changes in inclusion/exclusion criteria to an important variation in the final population. Another source of variability between studies is the sample size and the presence or absence of adequate cross-validation procedures, which greatly influence the results reported. For instance, Baumgartner et al. [24] use a population of 88 patients with a 50% design-validation split, while Varpula et al. [282] benefit from a larger sample ($n = 250$) but present *in-sample* analysis (no cross-validation).

3.3.2 Performance on the sepsis population

Some techniques and scores described in this chapter require covariates that are not directly available in the MIMIC-II database. For the others - essentially general severity scores - strong evidence has been presented earlier, suggesting that all scores (SAPS, APACHE, MPM) perform similarly in a population of patients with sepsis syndrome. Other reports demonstrate the equivalence of APACHE and SAPS on general ICU populations [152]. We chose to implement the APACHE score because it benefited from the larger sample size for design and it has been updated regularly. Other scores could also easily be implemented: SOFA, SAPS-I, and OASIS were also compared with the APACHE-IV score.

A three-fold cross-validation procedure was performed to evaluate performance. First, each score s was implemented as described in its original publication and calibration coefficients β_0 and β_1 were identified on the training set, so that $\hat{y} = \pi(\beta_0 + \beta_1 \cdot s)$. Calibration coefficients were then applied to the test set and estimated probabilities of death $\hat{y}$ were used to compute metrics of performance. Then, for each severity score, a new model was built using its initial covariates as inputs for a logistic regression model. For each covariate, the median values over the first 24 hours following admission were extracted for each patient. Observations departing from the mean of more than 3 standard deviations were discarded [33]. Each variable in the dataset was then normalized so that

Table 3.3: *Performance of different clinical scores of severity applied to a population of severe sepsis patients with treated hypotension: Acute Physiology and Chronic Health Evaluation (APACHE) III and IV, Sequential Organ Failure Assessment (SOFA), Oxford Acute Severity of Illness Score (OASIS), Acute Physiology Score (APS) and the Simplified Acute Physiology Score (SAPS). The results are extracted on three test folds and metrics are presented with mean and standard deviation: negative log-likelihood (Nll), Area Under the Receiver Operating Curve (AUROC), and Standardized Mortality Ratio (SMR). Model names with * indicates that all the co-variates included in the initial model were used as predictors rather than their scored value.*

Variable	Nll	AUROC	HL	HLp	SMR
APACHE-IV	385.8±3.8	71.1±0.4	8.6±5.0	0.065±0.031	0.90±0.08
SOFA*	387.5±4.9	70.8±1.4	17.3±3.6	0.014±0.017	0.93±0.07
OASIS*	387.7±8.1	71.4±0.8	15.1±2.4	0.021±0.012	0.82±0.06
APS*	389.9±13.8	71.4±2.3	24.3±7.9	0.006±0.009	0.82±0.08
APACHE-III	390.8±4.5	69.9±0.4	9.6±5.5	0.051±0.025	0.96±0.15
APS	393.8±7.3	69.3±1.3	6.1±2.0	0.100±0.011	1.00±0.03
APACHE-III*	397.9±37.6	73.4±2.2	9.6±4.5	0.061±0.033	0.91±0.09
APACHE-IV*	400.8±35.3	73.3±2.5	13.0±5.5	0.046±0.054	0.94±0.05
SOFA	402.3±11.0	65.7±2.9	13.4±3.4	0.037±0.034	0.90±0.03
SAPS-I	405.7±1.5	64.9±0.5	11.8±4.7	0.054±0.050	1.03±0.20
OASIS	417.0±3.3	61.4±2.6	11.2±7.5	0.072±0.059	0.98±0.17
VanWalraven	417.7±5.8	62.6±3.1	14.5±2.8	0.027±0.016	0.91±0.07
Elixhauser	424.7±20.3	64.0±1.2	20.9±12.8	0.037±0.064	0.77±0.02

$x_{normalized} = \frac{x-\mu}{\sigma}$ where μ and σ denote the mean and the standard deviation, respectively, computed from the training observations for the given variable. Variables that were visually found to have extreme tails were log-transformed as a rule of thumb. All covariates were then used to train a logistic regression model that was subsequently applied to the test set from which performance metrics were derived. At the end of this procedure, each performance metric was computed three times (once on each fold). Table 3.3 shows the performance metrics computed for each severity score on our population of patients.

3.3.3 Discussion

As expected, the different versions, of APACHE perform increasingly better on our population. The latest version of APACHE shows good calibration ($HL_{\hat{C}^*} = 8.6$, $p > 0.05$) but only offers an AUC of 71.2%. These findings confirm the results reported in the clinical literature and presented in the previous section: severity scores need customization and recalibration when applied to specific populations. Customization of the APACHE-III score using initial covariates and logistic regression improved model accuracy (AUC=73.5%) and attained good calibration ($HL_{\hat{C}^*} = 9.6$, $p > 0.05$). Conversely, the customization of APACHE-IV did not improve the calibration ($p > 0.05$) but also showed good discrimination (AUC=73.3%). This result suggests a disadvantageous ratio between the number of patients and variables, which stresses the need for more data or fewer covariates. Appropriate selection of covariates could advantageously be implemented with several techniques that we will presented in the next chapter.

In addition to this, the performance displayed by APACHE-IV seems to be on the lower side of expectations. Sepsis induces physiological derangements that are impacting the way physiology relates to the outcome. For instance, a patient with sepsis-induced hypotension and fluid resuscitation with normal blood pressure levels certainly has a higher risk of severe outcome than a similar patient who did not receive fluids. To illustrate, we show that physiology accounted for only 20.1% of the explained variance using the technique of χ^2 ratios described by [158, p. 1633]. This figures contrasts dramatically with the 65.7% reported on a more general population for APACHE-IV [313]. It has already been widely reported that the ability of a model to discriminate and calibrate decreases when applied to new populations [5] but this result suggests that these patients may be admitted to the ICU with a more homogeneous physiological derangement than a general ICU population.

Beyond physiological considerations, the implementation of APACHE-IV on the MIMIC-II database has limitations that are likely to alter its performance. For instance, the extraction of some covariates does not guarantee a high correlation with the way the APACHE

database was collected: comorbidities, mechanical ventilation, and the number of grafts used for cardiac surgery. In addition to this, some components were simply missing and consequently ignored during the score calculation including the admission diagnosis codes.

Most traditional severity scores were designed on data collected manually, sometimes through a web interface. The intervention of a human operator during the data collection process may provide some kind of benefit. For instance, physiologically impossible values could be corrected at the time of key-in and artefactual readings could be filtered out after a brief inspection of trends displayed by the information system. On the other hand, if no extra staff is hired to fill in the data, its quality would then undeniably be modulated by staff workload. For that reason and many others, such manual data procedure leaves much space for variations within and between individuals and hospitals. For instance, an understaffed institution would provide data of worst quality than a well-funded one (inter-hospital variability). Similarly, a clinician might collect data of worst quality on a day directly following a night shift. As a consequence, the rate of missing data may capture the hospital social background or the medical staff workload rather than anything else. This would lead to undesirable consequences on the data analysis [141]. Conversely, the use of an automated data collection system such as the one implemented in MIMIC-II potentially reduces the cost (secondary use of routinely collected clinical data) and the variability of the data. Naturally, such data is also prone to different types of noise requiring the implementation of alternative data pre-processing strategies. To illustrate, a MAP greater than 200mmHg would occasionally be found in the MIMIC-II database reflecting a pressure cuff attached to a bed handle; likewise a heart rate of zero corresponds to a detached ECG lead. Fortunately, the consistency of such artefacts across the entire database creates an opportunity for a common pre-processing strategy.

These results suggest that if customization presumably improves score performance, it is complicated by the small sample size. Alternative strategies have to be adopted to

deal with this issue. In particular, a strong effort on feature selection has to be made to avoid the drop in performance for observed scores with a large number of potential covariates.

3.4 Conclusion

The scores we presented here are based on a combination of variables that reflect pre-existing health as well as variables that reflect physiologic derangement due to acute illness. While these scoring systems have the potential to inform prognosis and resource allocation retrospectively at a cohort level in the ICU [276], their use has been mostly restricted to clinical trials, for case-mix determination in retrospective data analyses [310] and benchmarking ICU performance [4]. This is largely due to the fact that these scoring systems perform well in predicting outcome at the group level, but continue to perform poorly when predicting survival in individual patients. There are a number of reasons for the limited predictive ability of current systems. Pertinent causal factors such as genetic factors may be excluded [287]. Recently, it has been shown that identification of "worst" values over a day by clinicians is biased [141], which may partially contribute to lower prediction performance than should be possible. In addition to this, scores used to benchmark ICU performance can only use information that is not influenced by local practice (admission and following 24 hours) and therefore do not benefit from the potential prognostic value of later observations. Additional factors that are currently not well understood may also exist. Recent research has emphasized the importance of early goal directed therapy in reducing mortality from septic shock [243], increasing the need for accurate early warning systems. Changes in the physiologic variables measured within hours of this early critical period may be more predictive of outcome than focusing on the worst values measured within the day after ICU admission. While current scoring systems such as APACHE [313, 160, 158, 258] try to capture the severity of the initial insult, they are likely to be limited in their ability to capture the "physiologic reserve"

of a patient to respond to this insult because they tend to focus on the worst recorded value over 24 hours, and not on the variability in an individual's immediate response to the physiologic insult (for example hypotension) and its treatment. Additionally, current severity scoring systems have been developed using a knowledge-driven approach where predictors are chosen based on known clinical variables associated with poor outcome. Recent developments in the field of genetic epidemiology (the study of the role of genetic factors in determining health in populations) have demonstrated that study designs which use a heuristic and data-driven approach to select predictors [248] have the potential to discover new causal factors of disease. Such approaches are explored in the following chapters.

Chapter 4

Machine learning approach to prediction of mortality

4.1 Clinical data mining

4.1.1 A short introduction to pattern recognition

Artificial Intelligence (AI) can be defined as the "design of intelligent agents" [196, p.1] where "intelligence" can be seen as performance with respect to a specific task. Machine Learning is a field of AI that aims at designing algorithms, whose behaviour is derived from empirical data. In particular, Pattern Recognition (Pattern Recognition) techniques map input values to an estimated output with parameters derived from past occurrences. This task is also known as *modelling* and the function mapping the input values to an estimated output is often referred to as the *model*. Predicting mortality in a population of patients for which data and outcomes have previously been collected is a typical Pattern Recognition problem. More precisely, trying to map observations onto categories is called a *classification* task. Furthermore, when labels for the data are known, this problem is one of *supervised classification*, which is usually carried out in two steps:

Training the algorithm is provided with inputs (also called the "data", denoted as X, an N-by-P matrix where N is the number of patients and P the number of vari-

ables) and the outcome (also called the "target", denoted as y an 1-by-N binary vector). The model $\mathcal{F}$ has parameters θ that are fitted to minimize the error term $\epsilon = \hat{y} - y$ where the estimated output is $\hat{y} = \mathcal{F}(X)$; Sometimes the process of model fitting requires the estimation of the "out-of-sample performance", which is often achieved by splitting the training set into one or more smaller training and *validation* set(s) as detailed below;

Test unseen input data processed by the algorithm whose output ($\hat{y}$) is compared to the real outcome (y) allowing computation of performance metrics.

Because supervised Pattern Recognition techniques are meant to be applied to real-life problems, there are a few practicalities related to their design and use that will be discussed throughout this work. In fact, expertise in Machine Learning (Machine Learning) largely resides in know-how derived from hands-on experience.

Fitting is the procedure during which the optimal model parameters are estimated from the available data. The fitting procedure consists of minimizing an error term leading to the identification of the parameters as described by equation 4.1. Some Pattern Recognition algorithms have *analytic* solutions, which means that the optimal parameters can be directly written as a function of the data X. For most applications however analytic solutions do not exist and iterative techniques are used to minimize equation 4.1. These techniques, like gradient descent [227, p.420-425], usually assume convexity of the optimization criteria. In some cases however, such an assumption does not hold or the search space is too large, and other iterative techniques can be used like heuristics that trade the guarantee of finding the global optimum for convenience (speed, lack of hypothesis).

$$\theta = \underset{\theta}{\text{argmin}}(\mathcal{F}(X) - \hat{\mathbf{y}}) \tag{4.1}$$

Overfitting happens when a model describes (or fits) the existing data too well but poorly predicts new instances. An intuitive description of overfitting with illustrations is provided by Bishop [31, p33-36] and [32, p11-14]. There are two aspects of modelling that induce overfitting: (1) a sample size that is too small, and (2) a model that is too complex. If you have a dataset composed of two non-surviving male patients and one surviving female you may deduce that gender perfectly predicts mortality. Clearly, this is an erroneous inference generated by the small sample size. Figure 4.1 illustrates the second mechanism: let us assume we want to estimate the probability of someone being British based on his Global Positioning System (GPS) coordinates. In order to do so we randomly select a few dozen British men and record their GPS coordinates (black dots on the map). A simple approach would consist of drawing a straight line through the English channel (model A in orange) and saying that everybody north of it is British; a more complex model would consist of drawing a circle around Britain (model B in green) in order to exclude Nordic countries. Finally, a model that is too complex would draw a circle around each British citizen in the world (model C in red). The first model (A) is a typical example of underfitting since performance on the data used to train the model is non-optimal. On the contrary, the third model (C) provides excellent performance on this data but is likely to show a dramatic drop in performance on a data set drawn at a different time (grey points). The trade-off between model complexity and model fitting is related to Occam's razor [36] that support the choice of the simplest model (less hypothesis) at equivalent performance.

The curse of dimensionality The example presented above is a simple two dimensional problem with latitude and longitude as the input variables. In a clinical database however, and more generally in real-world applications, the number of observations and therefore the number of variables is potentially infinite. The three imaginary models that are represented in Figure 4.1 are increasingly complex and require 2 parameters for the line and 5 for the ellipse. If we now look for a prediction in 3 dimensions, models of equivalent complexity will now require the estimation of 3 and 7 parameters respectively. For problems

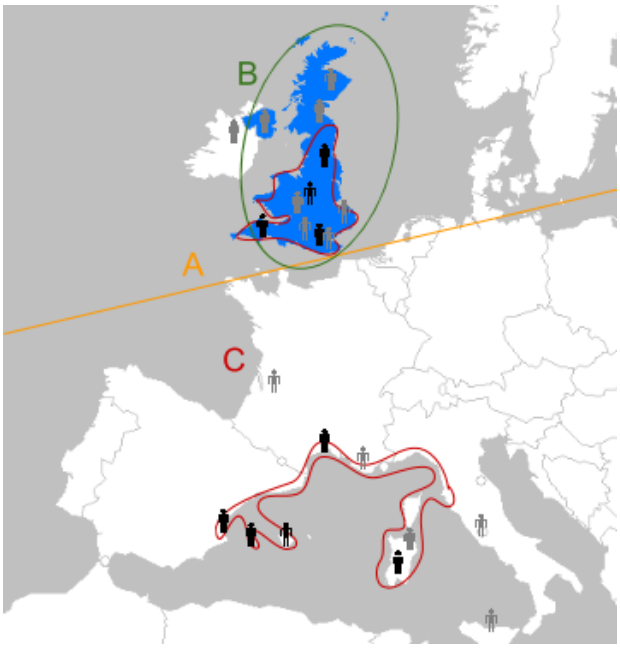


Figure 4.1: *Identification of British citizens based on GPS coordinates. Black points are random citizens selected to design the model. Three models of increasing complexity are designed: A, B and C. Grey points indicate the locations for a larger sample drawn from the population (maybe at a different time), which indicates that the model of adequate complexity shows better performance.*

of higher dimensionality, the number of parameters to be estimated is generally said to grow as a power of the number of dimensions [31, p. 33-37]. Unfortunately the number of observations available in practice is usually limited and the resolution of equations similar to 4.1 requires strategies to reduce dimensionality. This is usually achieved with feature selection [112] or variable transformation [168, p. 252-254]. Tackling the curse of dimensionality is one of the most important aspects of Machine Learning algorithm design and we will therefore discuss this issue further in this work.

4.1.2 Machine learning modelling techniques

We have introduced in section 3.1.1 the principle of LR, which has been used extensively to estimate the risk of death in the ICU. We introduce here more advanced techniques, namely Support Vector Machines (SVMs) and Random Forests (RFs).

4.1.2.1 Support vector machine (SVM)

A SVM is a non-probabilistic binary classifier. It tries to identify the best separating plane between observations $\mathbf{x_i}$ belonging to different classes. It first defines a family of hyperplanes

$$H_j(\mathbf{w_j}) : \mathbf{w_j} \cdot \mathbf{x} + b_j = 0 \tag{4.2}$$

which separates two classes in the hyperspace. The best separating plane $H_0 : \mathbf{w_0} \cdot \mathbf{x} + b_0 = 0$ is then taken to be the one that *maximizes* Euclidean distances to points for each class [280, 68], which corresponds to

$$\text{argmin} \quad \frac{1}{2} \|\mathbf{w}\|^2 \tag{4.3}$$

$$\text{subject to} \quad y_i \left(\mathbf{w} \cdot \mathbf{x_i} + b\right) \geq 1 \tag{4.4}$$

The reason why the hyperplane chosen maximises the margin is because it offers interesting properties. More precisely, it was demonstrated that the probability of misclassifying new observation (ϵ_{Test}) can be bound by the training error plus the *capacity*.

$$\begin{aligned} \epsilon_{Test} &\leq \epsilon_{Training} &&+\sqrt{\frac{1}{n} h \left(\log \left(\frac{2n}{h} + 1 \right) - log \frac{4}{\delta} \right)} \\ \text{Test Error} &\leq \text{Training Error} &&+\text{Capacity} \end{aligned} \tag{4.5}$$

The capacity is linked to the complexity of possible models and can be expressed as function of the Vapnik-Chervonenkis dimension (VC-dim) h [279], which in turn can be seen as the maximum number of points that can be separated in all possible ways by a set of functions (usually equal to the dimension of the space plus one). Interestingly, the VC-dim was proven to also be bound

$$h < \frac{R^2}{\rho^2} + 1 \tag{4.6}$$

where R is the radius of the smallest sphere containing all the data and ρ the margin (the distance between observations and the separating hyperplane). This essentially indirectly bounds the test error to a number that is independent of the dimension. For that reason, maximising the margin ρ shifts the optimisation problem from curse of dimensionality to that of capacity, which will prove particularly convenient when dealing with problems in high dimensions.

Solving this problem of margin maximisation defined by equations 4.3 and 4.4 can be achieved with the use of Lagrange multipliers $\alpha_i \geq 0$ and the Lagrangian function:

$$L(\mathbf{w}, b, \alpha) : -\alpha(4.4) + (4.3) \tag{4.7}$$

Minimize training error and Maximize margin

$$= -\sum_{i=1}^{\ell} \alpha_i \left[y_i \left(\mathbf{x_i} \cdot \mathbf{w} + b \right) - 1 \right] + \frac{1}{2} \mathbf{w}^2 \tag{4.8}$$

where $\Lambda^T = (\alpha_1, \ldots, \alpha_\ell)$ is the vector of non-negative Lagrange multipliers corresponding to constraints 4.4 and ℓ is the number of observations in the training set. It is worth noting that equation 4.8 reflects the bound on test error introduced in equation 4.5 and illustrates how capacity is contained by maximizing the margin. Solving equation 4.8 can be achieved with its derivative with respect to b and $\mathbf{w}$ [266, 280, 68] leading to

$$\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i y_i \mathbf{x_i} \tag{4.9}$$

$$0 = \sum_{i=1}^{\ell} \alpha_i t_i \tag{4.10}$$

However, in practice a hyperplane hardly ever perfectly separates classes. To account for this, a *soft*-margin can be used to introduce some flexibility (*slack*) in the formulation of the optimal hypersurface. This is described by Cortes and Vapnik [68] who reformu-

lated equation 4.3 as

$$\text{argmin} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{\ell}\xi_i \tag{4.11}$$

$$\text{subject to} \quad y_i\left[\mathbf{w}\cdot\mathbf{x_i}+b\right] \geq 1-\xi_i \tag{4.12}$$

$$\text{and} \quad \xi_i \geq 0, i=1,\ldots,\ell \tag{4.13}$$

where C is called the *slack* variable. When C tends to infinity, the decision boundary converges to the hard-margin solution. While C decreases the optimal hypersurface allows some misclassification of the training observations, which reduces the degree of overfitting by ignoring artefactual training observations. At the same time, a too small C value could oversimplify the separating hyperplane and lead to poor classification accuracy. Consequently, an optimal trade-off between training accuracy (complexity of the separating hyperplane) and generalization properties must be identified by adequately setting this hyper-parameter.

Conceptually however, it is difficult to represent exactly the meaning of variable C. An alternative formulation of the SVM classifier has therefore been introduced by Schölkopf et al. [251] who introduce a parameter ν corresponding to the proportion of authorized misclassified observations in the training data. As a result of this formulation, Chang and Lin [54] later demonstrated that

$$0 \leq \nu \leq 2\frac{\min(\aleph_{y_i=+1}, \aleph_{y_i=-1})}{\ell} \leq 1 \tag{4.14}$$

where $\aleph_{y_i=1}$ denotes the number of observations belonging to the group defined by $y_i = 1$. According to this equation, ν must be lower than twice the number of cases in the smallest class.

The classification of new observations will finally be performed with the following equation:

$$\hat{y} = \ sign\left(\sum_{i=1}^{\ell}\alpha_i y_i \mathbf{x_i} + b\right) \tag{4.15}$$

Interestingly, the optimisation of such problems requires the Karush-Kuhn-Tucker (KKT) conditions to hold, imposing for every data point either $\alpha_n = 0$ or $y_n \hat{y}_n \mathbf{x_n} = 1$ [31, Appendix E]. As a consequence, any data point for which $\alpha_n = 0$ will not contribute to the estimation of the prediction in equation 4.15. The remaining data points therefore satisfy $\hat{y}_n y_n \mathbf{x_n} = 1$ meaning that they lie on the maximum margin of the hyperplanes ($y_i \left(\mathbf{w} \cdot \mathbf{x_{sv}} + b\right) = 1$, where $y_i \in \{0, 1\}$). For that reason these observations were given the name of *support vectors* (sv). As we will see, the selection of a subset of training points for the prediction of new observations combines advantageously with the use of kernel techniques.

Kernel transformation An essential aspect of SVMs, which certainly has contributed to their popularity, is that data points that are not linearly separable can be transformed in a way that allows linear discrimination, which is known as the *kernel trick*. The techniques uses a non-linear mapping of the feature space $\Phi(\mathbf{x})$ and the kernel function is given by the following inner product of the feature space

$$k(\mathbf{x_i}, \mathbf{x_j}) = \Phi(\mathbf{x_i})^T \Phi(\mathbf{x_j}) \tag{4.16}$$

where $\mathbf{x_i}$ and $\mathbf{x_j}$ are two observations. The choice of the kernel function is potentially infinite as long as its Gramm matrix $K_{ij} = k(\mathbf{x_i}, \mathbf{x_j})$ (an $N \times N$ matrix) is positive semidefinite. We will restrict our analysis to the most common: polynomial kernels [182] and Radial Basis Function (RBF) kernels [206], which are described by equations 4.17 and 4.18, respectively.

$$Polynomial: \quad \Phi(\mathbf{x}) = \left(\sigma \mathbf{x}^T \mathbf{x} + c\right)^d \tag{4.17}$$

$$RBF: \quad \Phi(\mathbf{x}) = \exp\left(-\frac{||\mathbf{x} - \mathbf{x'}||^2}{2\sigma^2}\right) \tag{4.18}$$

Radial basis function kernels are by far the most popular in the machine learning literature possibly because of the advantageous trade-off they offer between performance and

complexity (number of parameters) [265]. Also, they are by design robust to outliers. Polynomial kernels offer equivalent performance but possess alternative properties [62].

The use of kernel function is not restricted to that of SVMs and can more generally be applied to any techniques where a *dual representation* can be formulated, in which the solution can be expressed directly in terms of the kernel function (K, the Gramm matrix). A first important results is that the kernel substitution breaks the straight relation between number of feature (model dimensionality), which seems particularly relevant to solve problems of high dimensionality. Yet, solving the problem with the dual representation requires the inversion of the $N \times N$ Gramm matrix rather than $M \times M$ matrix (number of variables) that can also be of much lower dimension (the number of observations is often larger that of features). The sparse solution provided by the SVM can be of particular use in this situation. In addition to this, working with the kernel function does not requires an *explicit* formulation of the non-linear mapping function Φ, which allows to work in a mapped feature space that is of a potentially infinite dimensionality. Using the dual representation, equation 4.15 becomes

$$\hat{y}(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{\ell} \alpha_i y_i k(\mathbf{x}, \mathbf{x_i}) + b\right) \tag{4.19}$$

where the threshold parameter is computed from N_{sv} support vectors as

$$b = \frac{1}{N_{sv}} \sum_{n=1}^{N_{sv}} \left(y_n - \sum_{m=1}^{N_{sv}} \alpha_m t_m k(\mathbf{x_n}, \mathbf{x_m})\right). \tag{4.20}$$

To conclude, SVMs offer a sparse solution to minimise the training error and maximise the margin between two classes in order to optimise the out-of-sample performance, which was found to be bound by capacity rather than dimension. In addition to this, the dual representation of this problem with the kernel function that does not require an explicit mapping function allows to non-linearly transform and separate the data in spaces of higher (and potentially infinite) dimension. This trick combined to the sparse solution offered by support vectors elegantly circumvent the problems of curse

of dimensionality. Finally, unlike other machine learning techniques such as Neural Network (NN), SVMs are a convex optimisation problem for which a single solution can be found.

4.1.2.2 Random forest

The Random Forest (RF) is another algorithm in machine learning. It belongs to the family of algorithms that use an ensemble of regression or classification trees. A decision tree is a structure in which each node is represented by a variable and a decision criterion. An observation can then be passed through the tree from the top node to the bottom: at each node, the observation value and decision criterion will define which branch to go to; ultimately, the final leaf (at the bottom of the tree) will indicate a value or a decision. There are many existing algorithms to build single decision trees such as ID3 [231], C4.5 [232], and CART (Classification and Regression Tree) [44]. The RF algorithm uses the CART technique that relies on the definition of a *Gini impurity* function:

$$I_G(n) = 1 - \sum_k p(k|n)^2 \tag{4.21}$$

where $p(k|n)$ is the probability of the class k at node n. When only two classes are present, the decomposition of $p(k = 0|t)^2$ reduces equation 4.21 to $I_G(n) = 2p_0(1 - p_0)$. This criterion is a maximum at $I_G = 0.5$ when classes are balanced and cancels out when the node t is pure (only positive or negative classes left). Each tree is *grown* by selecting at each node an optimal couple of variables and a cut-off value according to the maximization of a homogeneity" gain:

$$\Delta I_G(n) = I_G(n-1) - P_r.I_G(n_r) - P_l.I_G(n_l) \tag{4.22}$$

where $I_G(n)$ denotes the Gini impurity of the n^{th} node which splits into two branches (left and right) with associated probabilities P_l and P_r, respectively. The tree can be built fully, which means that homogeneity is reached at each leaf n_{max} ($I_G(n_{max}) = 0$).

Conversely, to prevent overfitting, the tree can be built only partly which means that a subset of the observations or of the variables will be used for each tree. For instance, the RF algorithm randomly builds each tree using N observations randomly selected *with replacement* and $m << M$ variables where N and M are the number of patients and variables, respectively.

In this setting, two parameters were shown by Breiman [44] to influence the error rate of RF: the *correlation* between any two trees in the forest and the *strength* of each individual tree. Increasing the parameter m was found to increase both the correlation and the strength of the tree. Consequently, this parameter constitutes the only important parameter for the algorithm. The presence of a single parameter to tune possibly accounts for the popularity of the approach.

4.1.3 Feature selection techniques

Another essential aspect of clinical data-mining is the ability to mine large number of observations and identify the variables that *matter* the most with respect to a specific outcome. This problem is closely related to feature selection techniques, which initially emerged to overcome the curse of dimensionality (introduced earlier in this chapter). Ultimately, the process of feature selection (1) allows the identification of the clinically relevant variables and (2) solves the curse of dimensionality in a context of limited data.

It is generally accepted that feature selection techniques fall into two distinct categories: *filter* and *wrapper* methods. Filter techniques first identify a number of covariates that are then passed-down to an independent classifier. A wrapper technique refers to an algorithm that automatically selects variables while deriving the model parameters from training data. These techniques seem conceptually more elegant and should be preferred since they identify the optimal features for the optimal classifier. However, to test and compare specific classifiers, a common filter approach can be chosen for all of them. Finally, not all feature selection techniques can be covered here and we choose a linear and non-linear state-of-the-art technique recommended by Tsanas et al. [274].

4.1.3.1 The least absolute shrinkage and selection operator (LASSO)

A straightforward way to solve the regression equation $\hat{y} = \beta_0 + \sum_{j=1}^{N} \beta_j x_j$, where x_j is the input vector of dimension K, is to minimize the Sum Square Error (SSE) of residuals which is called the Least Squares estimate (LSE) and is defined as follows:

$$\hat{\beta} = \text{argmin} \sum_{i=1}^{N} (y - \hat{y})^2 \tag{4.23}$$

Solving equation 4.23 can be achieved by setting the derivative of the regression equation to zero. This equation is called the *normal equation* (more details can be found in [309]). It involves inverting the matrix $\mathbf{x}^T\mathbf{x}$ which may be singular or nearly singular, in particular when the number of Events Per Variable (EPV) is too low or when some variables are correlated, leading to a high-variance unrealistic model. A first improvement was to suggest a stabilization or regularisation factor, so that $\mathbf{x}^T\mathbf{x} - \alpha I$, where α is a tuning parameter, would have to be inverted instead of $\mathbf{x}^T\mathbf{x}$. This was shown [132] to be equivalent to minimizing the following penalized loss-function:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{K} \beta_j^2 \tag{4.24}$$

This technique, also called *ridge regression*, can be seen as a LSE under the constraint that β should not be too large. It offers the advantage of stability but still has a major drawback: the coefficients shrink but are however never set to zero, which makes models difficult to interpret.

The Least Absolute Shrinkage And Selection Operator (LASSO) technique was later introduced in [270] and solves equation 4.23 subject to $\sum |\beta_j| < \alpha$, where $\alpha \geq 0$ is a tuning parameter. The LASSO can therefore be expressed as follow:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \alpha * |\beta| \tag{4.25}$$

When α is large, the LASSO provides a solution that is equivalent to that of the ridge

regression. When α tends to zero, some β_j coefficients eventually cancel, therefore discarding the feature that are irrelevant with respect to the LASSO criteria. It is a quadratic programming problem for which standard numerical analysis algorithms can be used to search for minima. Because of these advantages LASSO has gained popularity over the past decade and is a standard technique for feature selection which has been successfully applied to predict mortality [271].

Ridge regression and LASSO can be integrated in a larger framework that was earlier introduced in [94] by suggesting the use of a bound on the L^p-norm of the parameters, defined as $||\beta||_p = \sqrt[p]{\sum_i \beta_i^p}$, such that equation 4.23 can be regularized with L^p-norm as:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \sum (y - \hat{y})^2 + \alpha * L^p(\beta) \tag{4.26}$$

Ridge regression and LASSO can be seen as the L^2 and L^1 regularization of the LSE, respectively.

4.1.3.2 The Relief algorithm

The *Relief* algorithm is a popular feature selection technique [156] that offers an extremely advantageous trade-off between simplicity and performance. This point constitutes a very strong rationale for its use instead of more complicated techniques. For instance, the Minimum Redundancy Maximum Relevance (mRMR) technique [222] is inspired by Shannon and Weaver's [1948] Mathematical Theory of Communication (MTC) and has shown good performance [274] but comes with various implementation tricks, which are all potential sources of variability.

The pseudo-code of Relief is presented in Algorithm 4.1.1, which illustrates the simplicity of the approach. The technique takes three arguments: a data set $\mathcal{S}$, a number of iterations m, and an integer κ. The latter indicates the number of closest neighbours to be considered for the *random* selection of near hits and misses at lines 1 and 2. The relevance metric for feature x can therefore be seen as the probability of observing a

different value of x given the nearest instance from a different class minus the same probability given the nearest instance from the same class, which are computed over all observations from the training set.

Algorithm 4.1.1: Refief($\mathcal{S}, m, \kappa$)

---

main

Separate $\mathcal{S}$ into $\mathcal{S}^+$ and $\mathcal{S}^-$

for $i \leftarrow 1$ **to** m

do
- Select a random $X \in \mathcal{S}$
- Select a random $Z^+ \in \mathcal{S}^+$ close to $(\kappa)X$
- Select a random $Z^- \in \mathcal{S}^-$ close to $(\kappa)X$
- **if** $X \in \mathcal{S}^+$
  - **then**
    - Near-hit $= Z^+$
    - Near-miss $= Z^-$
  - **else**
    - Near-hit $= Z^-$
    - Near-miss $= Z^+$
- **for** $j \leftarrow 1$ **to** p
  - **do** $W_i = W_i - \text{diff}(x_i, \text{Near-hit}_i)^2$ $\ldots + \text{diff}(x_i, \text{Near-miss}_i)^2$ (3)

$Relevance \leftarrow W/m$

return $(Relevance)$

This simple implementation however does not offer the possibility of dealing with missing values. *ReliefF* (note the extra "f") is an updated version of the algorithm proposed by Kononenko et al. [164] who suggested that the *diff* function used on line 3 of Relief (see Algorithm 4.1.1). The update switches to a probabilistic estimation should

either of the two values for the attribute A be missing :

$$\begin{aligned} \mathbf{x_i} \text{ is missing} \qquad & \text{diff}(\mathbf{x_i}, \mathbf{x_j}) = 1 - P\left(\text{value}(A, \mathbf{x_j}) | \text{class}(\mathbf{x_i})\right) \\ \mathbf{x_i}, \mathbf{x_j} \text{ are missing} \qquad & \text{diff}(\mathbf{x_i}, \mathbf{x_j}) = 1 - \sum_{V}^{\text{values(A)}} \left[P(V|\text{class}(\mathbf{x_i})) \times P(V|\text{class}(\mathbf{x_j}))\right] \end{aligned}$$

4.2 State-of-the art modelling for the prediction of mortality: The Physionet challenge

In 2012, the Physionet [107] network organized the challenge of predicting mortality [143] based on the MIMIC-II database. The dataset was composed of 12,000 randomly selected patients aged over 16 for whom at least 48 hours of data was available. For this general ICU population, 41 variables were extracted over the first two days from admission, and 36 of these were physiological variables. The outcome to be predicted was chosen to be in-hospital mortality and the scoring criteria were :

Score 1 was defined as the minimum value of Sensitivity (Se) and PPV that are described in section 3.1.2.3: the maximum for this value was chosen to provide a reasonable trade-off between discrimination and prognostic value in the context of a highly-skewed class distribution;

Score 2 was a modified Hosmer-Lemeshow statistic $HL_{\hat{C}}$ (described in section 3.1.2.2) $Score2 = \frac{H}{\pi_{10} - \pi 1}$ where π_i is the mean estimated risk for patients falling in the i^{th} decile of predicted severity. This measure was used to assess model calibration.

However, Johnson et al. [145] demonstrated that negative log $\mathcal{L}$ offered the advantage of reflecting reflect model calibration and discriminatory power while showing an overall lower variability. This finding is in line with the results we presented in section 3.1.2. Consequently, it was suggested that this metric should be used instead of Score 1 and 2.

The Physionet challenge has obvious links with the work presented here (same database, outcome, and equivalent variables) and therefore deserves special attention. Hence, the models performing best on each score are detailed in this section, implemented, and applied to our data. Table 4.1 shows the results of the top two entries for each score. In this table, the model submitted by Johnson et al. [145] was sub-optimal; the corrected version – submitted after the challenge – did rank first in both Score 1 and 2.

Table 4.1: *Scores of the top two performers in the Physionet challenge 2012 showing scores and respective ranks in the competition. The ranks indicated with a $^{(*)}$ are not official ranks but results from submission given after the challenge deadline.*

Name	Score 1 (rank)	Score 2 (rank)
Bayesian Ensemble (corrected)	0.5352 ($1^{st(*)}$)	13.67 ($1^{st(*)}$)
Cascaded SVM-GLM paradigm [62]	0.5353 (1^{st})	29.86 (5^{th})
Bayesian Ensemble [145]	0.5345 (2^{nd})	17.88 (1^{st})

4.2.1 Bayesian Ensemble of forests

The Bayesian ensemble of forests is an original idea of Dunkley et al. [145], which performed remarkably well during the Physionet Challenge. It belongs to the category of techniques that uses a forest of decision trees and similarly offers interesting properties such as variable importance and an elegant way to handle missing values.

Pre-processing One advantage of this technique is that little pre-processing is required before growing the forest. Indeed, variables are simply converted into ranks or quantiles of a normal distribution. The parameters required to transform the data are derived from the training data, saved, and subsequently applied to new observations (the validation set).

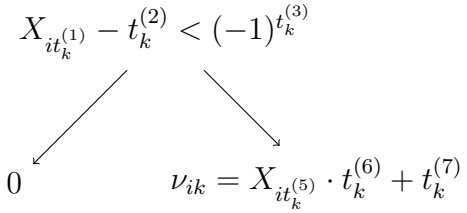


Figure 4.2: *Simple schematic of a tree. At the top node, variable $k^{(1)}$ is selected for the i^{th} patient giving value $X_{it_k^{(1)}}$. A split criteria is defined with $t_k^{(2)}$ and $t_k^{(3)}$. On one side value ν_{ik} for i^{th} patient and k^{th} tree is set to null. On the other branch, a second value $X_{it_k^{(5)}}$ taken on the same patient from variable $t_k^{(5)}$ is transformed with intercept $t_k^{(7)}$ and regression slope $t_k^{(6)}$ to give the final value of the tree ν_{ik} for patient i. on k^{th}*

Forest structure The forest $\mathcal{F}$ is composed of k trees and two additional parameters β_0 the intercept and β_1, the width, which will be described later on. The typical structure of a tree is represented in Figure 4.2, which introduces seven parameters t_k that determine ν_k, the output of the tree. The tree is defined by a split criterion $s_{ik} = X_{it_k^{(1)}} - t_k^{(2)} < (-1)^{t_k^{(3)}}$ and a regression rule (on a different co-variate) $X_{it_k^{(5)}} \cdot t_k^{(6)} + t_k^{(7)}$ that defines the tree output. Two additional parameters $t_k^{(4)}$ and $t_k^{(8)}$ replace potential missing values in $X_{it_k^{(1)}}$ and $X_{it_k^{(5)}}$, respectively. Once the forest is grown, the estimated output was defined with the logistic function (see section 3.1.1) as follow

$$\hat{\mathbf{y}} = \mathcal{F}(X) = \frac{1}{1 + \exp^{-\left(\beta_0 + \beta_1 \sum_{k=1}^{K} \nu_k\right)}} \tag{4.27}$$

Initialization The forest is initialized with $n_{tree} = 500$. Each parameter $t^{(j)}$ is randomly initiated from a prior distribution: $t^{(1)}$ and $t^{(5)}$ are sampled from a distribution with flat prior between 1 and P, the number of variables; other parameters are sampled from a standard normal distribution. β_0 is set as the prevalence of the training data and β_1, the width, is initialized as

$$\beta_1 = \frac{2 * \sqrt{\text{var}(\frac{1}{1+e^{-\hat{\mathbf{y}}}})}}{N_{tree}} \tag{4.28}$$

after deriving $\hat{y}$ from a simple model such as regularized logistic regression.

Growing the forest After initialization, the forest is updated stochastically with a Markov Chain Monte Carlo (MCMC) method as follows. At each iteration the posterior of the parameters given the data are given by:

$$p(t|X) \propto p(X|t) \times p(t) \tag{4.29}$$

where $p(X|t)$ and $p(t)$ denote the likelihood and priors, respectively. Each tree in the forest is replaced by a tree sampled from the prior distribution with a probability of 0.5 and otherwise left unchanged. As a consequence, the new forest depends only upon the current forest, which makes the iterative process essentially "memoryless". The new forest $\mathcal{F}_{new}$ is kept if it has a probability of generating the data $\mathcal{L}_{new}$ that is higher than the previous forest ($\mathcal{L}_{old}$). Otherwise, it is kept with a probability equal to the likelihood ratio, which corresponds to the following acceptance criterion:

$$\frac{\mathcal{L}(\mathcal{F}_{new})}{\mathcal{L}(\mathcal{F}_{old})} \geq \mathcal{N}(0,1) \tag{4.30}$$

The forest is initially left to burn during $N_{burn-in}$ iterations and then saved after every N_{save} iteration until N_{max} is reached. This process can be repeated N_{rep} sequentially or in parallel to speed up the process. Eventually, all of the saved forests are merged into a single giant forest of size $N_{tree} \times N_{rep} \times \mathcal{I}\left(\frac{(N_{max}-N_{burn-in})}{N_{save}}\right)$ where $\mathcal{I}$ denotes the integer part of a real number.

4.2.2 Cascaded SVM-GLM paradigm

The entry ranking second in the Physionet challenge was a cascaded SVM-GLM paradigm [62], which can be decomposed in three levels. First, each variable is mapped to a Gaussian distribution; then balanced data sets of transformed data are extracted to train different models (SVM); finally, the outputs are used to train a logistic regression model.

Pre-processing Minimum, maximum and median values were extracted from time series over two time intervals: the first 24 hours from admission and the following 24 hours (i.e. from 25 to 48 hours from admission). Categorical variables were transformed into $k-1$ binary variables where k is the number of categories. In addition to this, the sum of all urine records was taken as a new feature.

Variable normalization Citi and Barbieri [62] transformed each input vector $\mathbf{x_j}$ by a combination of a constant, the variable and the log of it using the assumptions that this procedure could successfully achieve normalize the data. All continuous variables were transformed to fit a normal-like distribution. Each observation x_{ij} was transformed into the corresponding percentile q_{ij} for the j^{th} variable. Then, regression coefficients b_j were identified so that

$$\Phi^{-1}(\mathbf{q_{ij}}) = (\mathbf{b_j})\mathbf{R^T} \tag{4.31}$$

where $R = \frac{1}{3}[\overrightarrow{1}, \mathbf{x_j}, \log(1+\mathbf{x_j})]$ is a $N \times 3$ matrix, $\overrightarrow{1}$ denotes a vector of ones, and $\Phi^{-1}(\mathbf{q_{ij}})$ is the i^{th} percentile of the inverse Cumulative Distribution Function (CDF) of a standard Gaussian distribution for variable j. In order to minimize the influence of outliers, only the observations meeting $q_i > 1$ and $q_i < 99$ were used to compute $\mathbf{b}$. Finally, the coefficients $\mathbf{w_j} = (R^T R)^{-1}(R^T \Phi^{-1}(\mathbf{q_{ij}}))$ were used to transform each variable as: $\mathbf{x'_j} = \mathbf{w_j} R$. Citi and Barbieri [62] demonstrate that this procedure successfully transforms variables towards a normal-like distribution.

Classification The transformed data was then used to train an ensemble of SVMs for classification. In particular, to account for the unbalanced nature of the dataset (mortality rate of 14.2% during the Physionet challenge) different models were built, each using all positive occurrences in the data and an equivalent number of negative cases sampled without replacement. In the end, six $(100/14.2)$ different ν-SVMs were trained.

Voting Each training observation was then passed down these six classifiers and a logistic regression model was trained to combine their output into a probability of risk.

Ideally however, and to prevent over-fitting during the training session, the training set should be split into three distinct parts in order to independently estimate the parameters for normalization, modelling and the estimation of the probability of risk.

4.3 Implementations on the severe sepsis population

4.3.1 The data set

The results presented in the previous chapter suggest that the covariates included in the APACHE-IV model are the most probable given the data (likelihood), which is in line with the available literature [16]. We have therefore implemented the different modelling techniques introduced in this chapter using covariates included in the APACHE-IV model. All of the variables that were not collected at admission (demographics and administrative variables) were aggregated over the first 24 hours with minimum, maximum, and median functions resulting in three features per covariate.

Outliers were simply identified using knowledge-based rules and a combination of percentiles and standard deviation as described in section 2.4. In addition to these, artefactual values were defined as being any univariate observation departing by more than 4 standard deviations from the mean, which were computed from the training data. Artefactual and missing values were simply imputed with the mean derived from the training set.

The performance of each technique was evaluated during a 3-fold validation procedure using the same split as in the previous chapters. Finally, all variables were normalized as

$$\mathbf{x}_{\mathbf{norm}} = \frac{\mathbf{x}_{\mathbf{old}} - \mu}{\sigma} \tag{4.32}$$

where μ and σ denote the mean and standard deviation derived from the training set, respectively. This normalization procedure was not implemented on the cascaded SVM as it already includes a normalization procedure.

4.3.1.1 LASSO

The LASSO model requires the identification of the optimal α value in equation 4.24. This was achieved within a 5-fold cross-validation procedure during which a geometric sequence of α providing non-null models was tried. This cross-validation level (local) has to be distinguished from the higher-level cross-validation procedure (global) that is three-fold. At each global fold (out of three), the value reducing the model deviance on the five local folds was saved and the LASSO was solved for the entire training set.

4.3.1.2 Grid search and filter feature selection for SVM

The implementation of an efficient SVM classifier requires the identification of the types and parameters of the kernel and SVM structure (slack variable or ν, width of the kernel or dimension). In addition to this, the SVM framework does not support an elegant wrapper feature selection technique. Consequently, we first identified the most important variables with the help of the ReliefF algorithm and then grid searched all of the available parameters for the different types of SVM and kernel we introduced above.

At each fold, the ReliefF algorithm was used to rank features according to their relevance (defined in line 4 of algorithm 4.1.1). Then for all deciles of input features ($0 : 0.1 : 1$), all possible combinations of kernel (restricting analysis to linear, polynomial, and RBF kernels) and SVM (ν and C) were explored (6 models) using a simple (and time-consuming) grid-search approach for the identification of the model's parameters. For C-SVM, the parameter C was taken from $2^{-11:3:1}$; for Nu-SVM, the parameter ν was taken at five equally spaced points between 1 and ν_{max} as defined in equation 4.14. For RBF kernels, the σ was sampled at $2^{-11:3:1}$; for polynomial kernels, the offset was set to $c = 0$ and the degree to $d = 3$ by default. The combination of parameters minimizing the negative log-likelihood was then searched with a second refined grid to identify optimal parameters. At each model fitting, the time required to fit the model was saved in order to estimate the computational cost of different SVMs and kernels with respect to the parameters' values.

4.3.1.3 Cascaded SVMs

A significant difference between this work and the Physionet challenge is the number of observations available to fit the model: only a sixth of the observations are left for training in this work. In their original article, Citi and Barbieri [62] fitted six different SVMs with 108 variables to balanced groups of roughly $1,000$ patients. The dataset presented here has smaller dimensions and requires the prior identification of features of interest. This will be achieved similarly to what has been presented in the previous section: the model will be run ten times, each time including an additional tenth of the top input features (ranked by relevance). Then, because the mortality rate is three times as big in this population, only three balanced datasets can be used to fit different SVMs. The outputs of all three SVMs are used as inputs for a logistic regression model fitted to the same training data.

4.3.2 Results

The results of this experiment are presented in table 4.2 that shows the different models (column 1 and 2) and their performance estimated on the validation sets during the three-fold cross-validation (columns 3 to 6). The LASSO model offers the smaller negative log-likelihood (Nll $= 369.4 \pm 5.9$, AUROC $= 75.05\%$), while the Bayesian Ensemble of Forests (BEF) ranks third after the RF and offers the highest metrics of discrimination (AUROC $= 76.16\%$). In fact the very The difference in AUCs was not found to be statistically significant ($p = 0.12$). Of all the SVM models, the C-SVM model consistently provides better performance on this dataset. Similarly, the linear and RBF kernels outperform the polynomial formulation. Finally, the cascaded-SVM did not provide good results.

Table 4.3 shows the parameters selected at each fold for the different combinations of SVM and kernel types. In addition to the raw performance presented in table 4.2, this indicates the stability of parameter identification over different folds. Figure 4.3 shows the range of parameters and the performance reported during the second-level five-fold cross-validation procedure. It indicates that all parameters remain stable from one fold

Table 4.2: *Performance of different modelling techniques for prediction of mortality in the severe sepsis patients with treated hypotension: Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RF), Bayesian Ensemble of Forests (BEF), C- and Nu-Support Vector Machine (SVM) with linear, polynomial and Radial Basis Function (RBF) kernels. The results are extracted using three test folds and aggregated with mean and standard deviation; performance metrics are: negative log-likelihood (Nll), Area Under the Receiver Operating Curve (AUROC), and Standardized Mortality Ratio (SMR).*

Models	Kernel	Nll	AUROC	SMR
LASSO	N.A.	369.4±5.9	75.0±0.8	0.9±0.1
RF	N.A.	371.2±3.4	74.3±0.5	0.9±0.1
BEF	N.A.	374.4±1.1	76.1±0.2	0.9±0.1
C-SVM	Linear	376.1±5.3	74.2±1.3	0.9±0.1
C-SVM	RBF	377.0±3.4	74.3±0.8	0.9±0.0
Nu-SVM	RBF	385.4±2.6	72.5±1.5	0.9±0.0
Nu-SVM	Linear	393.7±13.7	72.1±1.6	0.9±0.1
C-SVM	Polynomial	415.0±7.2	66.9±5.1	0.9±0.1
CascadSVM	N.A.	418.6±12.4	71.1±1.2	0.9±0.0
Nu-SVM	Polynomial	419.1±17.0	68.0±1.8	0.9±0.1

to another, with respect to the number of variables included in the model.

Finally, figure 4.3 (TOP) presents the computational cost of SVM fitting for different types of SVM and kernels and across different values of parameters. The lower part of the figure shows the cross-validated performance on the training set. Finally, table 4.3 shows that most APACHE-IV covariates are consistently selected to be included in the SVM model. The analysis of non-null weights given to different variables in the LASSO model is given in table 4.4.

4.4 Discussion

In this chapter, we have presented tools that seek to improve upon the traditional approach toward the selection of relevant clinical covariates (expert-driven) and linear modelling techniques to estimate the risk of death in the ICU. We described some techniques for feature selection (LASSO and ReliefF) and classification (RF and SVM). Some

Table 4.3: *List of parameters identified for each fold for C- and Nu-SVMs using linear, polynomial and Radial Basis Function (RBF) kernels. The last column shows the negative log-likelihood on the test set.*

SVM and Kernel Type	Fold	Parameters	Features Number (Percent)	N-ll
C-SVM (Linear)	1	$C = 2^{-3.998}$	45 (100.0)	377.8
	2	$C = 2^{-3.000}$	33 (73.3)	368.4
	3	$C = 2^{0.000}$	27 (60.0)	373.0
Nu-SVM (Linear)	1	Nu = 0.55	38 (84.4)	376.7
	2	Nu = 0.54	44 (97.8)	378.2
	3	Nu = 0.54	23 (51.1)	399.2
Nu-SVM (Polynomial)	1	Nu = 0.49, $\sigma = 2^{0.004}$	43 (95.6)	396.9
	2	Nu = 0.48, $\sigma = 2^{2.000}$	45 (100.0)	417.8
	3	Nu = 0.44, $\sigma = 2^{2.000}$	39 (86.7)	406.9
C-SVM (Polynomial)	1	$C = 2^{0.438}$, $\sigma = 2^{-6.000}$	36 (80.0)	394.2
	2	$C = 2^{-3.000}$, $\sigma = 2^{-4.969}$	31 (68.9)	433.3
	3	$C = 2^{-6.008}$, $\sigma = 2^{-7.250}$	9 (20.0)	432.2
Nu-SVM (RBF)	1	Nu = 0.49, $\sigma = 2^{-7.000}$	45 (100.0)	381.5
	2	Nu = 0.48, $\sigma = 2^{-4.492}$	44 (97.8)	393.4
	3	Nu = 0.48, $\sigma = 2^{-6.516}$	45 (100.0)	390.9
C-SVM (RBF)	1	$C = 2^{1.000}$, $\sigma = 2^{-9.563}$	45 (100.0)	377.8
	2	$C = 2^{-3.502}$, $\sigma = 2^{-13.000}$	41 (91.1)	372.5
	3	$C = 2^{1.000}$, $\sigma = 2^{-7.125}$	30 (66.7)	385.0

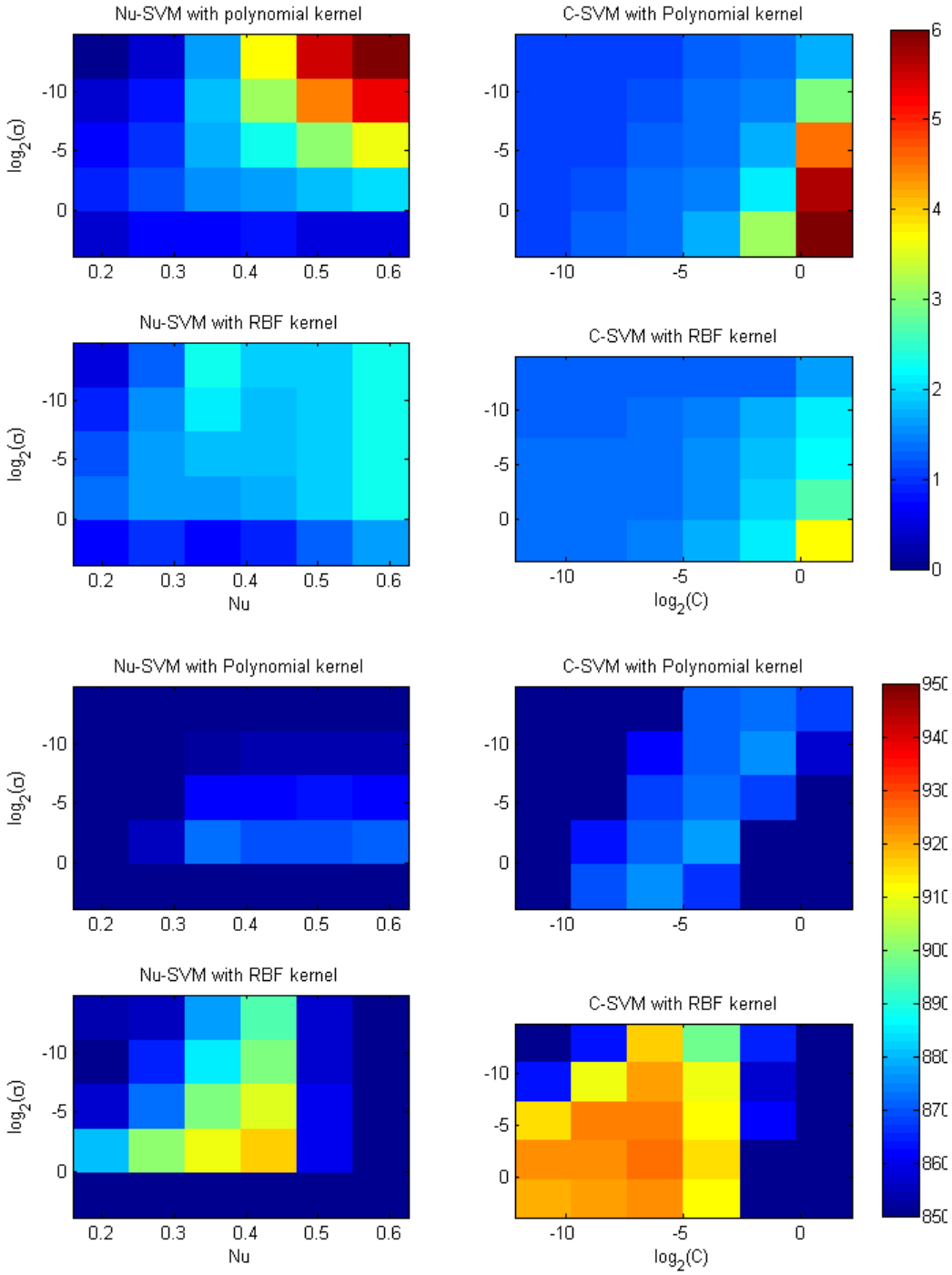


Figure 4.3: *(TOP) Log of the average time (in seconds) required to fit an SVM to the training data ($n \approx 1,440$ with about 20% positive cases). Lines one and two show the results for the polynomial and Radial Basis Function (RBF) kernels, respectively. Columns one and two show the Nu- and C-SVMs, respectively. Each cell is a colour-coded representation of the log of the time required for training the SVM given the parameters indicated in the abscissa (ν or log(C)) and ordinate (σ). The colour bar on the right-hand side of the picture indicates the numerical correspondence. (BOTTOM) Average negative log-likelihood estimated from different validation sets (five-fold validation) for the identification of the parameters (first level grid-search).*

Table 4.4: *Non-null variables selected by LASSO showing statistically significant coefficients ($p < 0.05$) sorted by decreasing absolute value of βs. The second column indicates whether the value was recorded before or at admission aas well as the aggregate function (minimum, maximum, mean) used to represent the samples collected during the first 24 hours.*

Variable	Aggregate	Value
APS	Mean	0.71
GCS	Mean	-0.30
Elective	Admission	-0.27
Age	Admission	0.26
Metastatic Cancer	Admission	0.19
PaFiO_2 Ratio	Max	-0.17
Surgery	Mean	-0.14
Lymphoma	Admission	0.10
Pre-LOS	Admission	0.09
Graft Number	Admission	0.07
Creatinine	Min	0.05
Emergency	Admission	0.05
AIDS	Admission	-0.04
GENDER	Mean	0.03
Rikers scale	Admission	0.03
Admission is transfer	Admission	-0.02
Bypass surgery	Admission	-0.02

Abbreviations: Acute Physiology Score (APS), Glasgow Coma Scale (GCS), Length of Stay (LOS), Acute Immune Dysfunction Syndrome (AIDS).

attempts have been made to empirically compare different machine learning algorithms [51] showing superiority of some groups of techniques including those we present here. Beyond these trends, the advantage of one technique over another is very much dataset-dependent including factors such as the number of observations and features, the nature of the relation between them, the quantity and randomness of missing data, and the presence of artefacts. Other mainstream classification techniques, such as neural networks and naïve Bayes, are not presented in this section. Logistic regression has been applied widely to estimate the risk of mortality rather than naïve Bayes and has proven to offer equivalent performance [246]. Similarly, recent approaches have successfully implemented SVM [62] and decision trees [145] for the prediction of mortality. Because not all existing machine learning techniques could be presented and implemented in this work we have decided to restrict our analysis to these described in this chapter.

Interestingly, the results presented in table 4.2 suggest that simpler models perform better. Indeed, LASSO does as well or better than any other model and the linear kernel compares advantageously to non-linear ones. While such results seem to contradict comon sense, we have found earlier reports of models like RF providing worse results than logistic regression [102]. With regard to the dataset used here, the limited amount of available observations left for training within the cross-validation procedure ($n \approx 1,440$) certainly accounts for the apparent superiority of simpler approaches. This phenomenon is even stronger when additional levels of cross-validation are required to identify parameters (such as kernel coefficients) on a training set. However, for larger datasets, sophisticated machine learning techniques [62, 145] clearly outperform logistic regression as seen during the last Physionet challenge. This however comes at a cost, in that the complexity may dramatically raise the threshold of acceptance amongst clinicians. Interestingly, all recent updates of the severity scores used logistic regression rather than more complicated models [313, 130, 131, 145].

Equivalent changes in a physiological variable does not always relate to the same change in severity. In particular for normally distributed *physiological* variables, the risk

is not necessarily symmetric with respect to the mean value; for instance an extremely low blood glucose value is certainly associated with a higher risk of dying than a high value that is equally departing from the mean. Clearly, a logistic regression framework is not expected to account for these and would equally penalize this two derangements. Conversely, a classification tree where each node represent a threshold on a given variable could adequately handle such non-linear behaviour. Interestingly, the comparison between different techniques suggests that the use of a non-linear machine learning technique meant to handle such property of the data might not necessarily be of primary importance. Indeed the linear logistic regression framework offered equivalent performance to non-linear techniques such as RF. The theoretical argument is tempered by the practical implementation of these techniques on this dataset of limited size.

To conclude, more sophisticated techniques should be preferred whenever the sample size is large enough to show a statistically significant improvement in performance; then, depending on the context, the gain in performance may or may not justify the increase in complexity. In particular, an algorithm that is fully integrated with the hospital information system could produce a simple probability and hide its real complexity from the end-user. We will touch on the relation between complexity and performance later on in this work.

Contrasting computational cost and performance as presented in figure 4.3 leads to additional interesting comments. First of all, larger values of ν for the Nu-SVM and larger values of C for the C-SVM lead to larger computational costs. These also correlate with a drop in performance on the training set: a larger regularization coefficient C in 4.11 and an increased proportion of misclassified training occurrences (ν). Likewise smaller values of σ for the polynomial kernel (equation 4.17) and larger values of σ for the RBF kernel (equation 4.18) lead to a significant increase in computational cost. Consequently, for extreme values of the couples (ν, σ) and (C, σ) the training time of a single SVM on roughly $1,500$ observations can be as long as ten minutes and increase exponentially. Nonetheless these computationally costly areas of the search-space are not those asso-

ciated with higher performance. Consequently, a heuristic trying to identify the optimal parameters for these models should avoid these costly areas and restrict the search-space to more relevant values of the parameters. Last but not least, the C-SVM offers both higher performance and lower computational cost, which is also true of the RBF kernel. This combination of SVM and kernel type should certainly be a preferred choice for an SVM architecture.

The low performance of the cascaded-SVM ($Nll = 419.1$, $AUROC = 68.0\%$) was not generated by a phenomenon of overfitting since the training results were found to be equally disappointing. Instead the different parameters that were set arbitrarily in the initial publication [62] may be responsible for this: Nu-SVM of $\nu = 0.5$, polynomial kernel with $d = 2$, $c = 0$, and $\sigma = 2^{\{-2,0,1\}}$. Indeed, the results presented in tables 4.2 and 4.3 suggest that neither the Nu-SVM nor the polynomial kernel look like a good first choice. In addition to this, the table suggests that the optimal values for the parameters do not match those suggested by Citi and Barbieri [62]. Yet the good performance reported in the Physionet challenge reinforces the confidence in the approach. The process leading to the identification of the correct parameters was unfortunately not described by Citi and Barbieri [62]. Using the results presented in this section to improve the performance of this model would certainly make use of the information derived from the validation set and generate a bias that may later impair the generalization properties of the technique. Alternative techniques need to be identified to identify precisely the variables of interest and the right parameters for a combination of SVM and kernel type.

Finally, for the estimation of variable importance, there exists an alternative implementation of the SVM with hyper-parameters on each of the input variables, which can indicate the relative contribution of each of the co-variates [272]. Such an approach however sometimes comes at the cost of inferior performance while we have alternative ways of reporting variable importance. The list of variables and statistically significant weights provided in table 4.4 highlights a few interesting facts. First of all, major contributions to the model are achieved by APS, GCS and age, which corroborates the findings

of most studies. Then, admission information that was included in most recent severity models also contribute equally: elective and post-surgery admissions are associated with decreased severity. This could be explained by three factors: first, the relative importance of post-cardiac surgery patients at BIDMC; second, the overall lower severity of these patients (ICU mortality of 2.6% as seen in table 2.2); and third, the fact that these patients may be screened more carefully for potential infection resulting in earlier management of the conditions and therefore lower severity. Finally, CHC and in particular metastatic cancer, lymphoma and AIDS were all found to contribute to severity, which also corroborates the results of earlier studies. The negative contribution of AIDS may have to been due to the different inclusion rule for the ICU. All together these findings on the contribution of variables are perfectly in line with the existing literature on the topic which certainly strengthens the confidence we have in the dataset.

4.5 Conclusion

In this chapter, we briefly introduced some of the key concepts of machine learning and described the theoretical framework behind techniques that go beyond the traditional logistic regression model. In particular, non-linear modelling techniques were described in the hope that they would capture non-linearities in the data and translate to an increase in performance. SVMs are maximum-margin classifiers designed to identify the best separating hyperplane between different classes. Different implementations exist such as Nu and C-SVMs that can be used in addition to non-linear data operators called kernels (linear, RBF kernel, and polynomial). The RF on the other hand is a non-linear ensemble of weak decision trees offering interesting properties including the ability to account for a large number of input variables.

In the data-rich environment of the ICU where thousands of variables are recorded for each patient, the need for feature selection is driven by two factors: to address the curse of dimensionality, and to provide parsimonious models that are easily interpreted,

for which acceptance amongst clinicians is likely to be higher. LASSO is a wrapper feature selection technique for logistic regression. ReliefF is a filter feature selection technique that computes a relevance metric for each variable according to the univariate distances of the nearest neighbours in each class. Finally, the recent Physionet challenge for the prediction of mortality on the MIMIC-II database featured models that were inspired by recently proposed techniques. The winning algorithm was a Bayesian ensemble of forests, similar to some extent to the random forest. The second ranked entry was a cascade of SVMs with a logistic regression model.

The most important result in this section is the significant improvement provided by the use of machine learning techniques in the place of simple logistic regression. In fact, compared to the results presented in the previous chapter, machine learning techniques provide an increase in negative log-likelihood from 385.8 to 369.4 and in AUC of 73.3% to 75.05%, which were both found to constitute a statistically significant improvement ($p < 0.05$). This improvement is primarily brought by variable pruning. The analysis of the results obtained with the techniques described in this chapter promotes a strong rationale for the use of simpler techniques. This is possibly imposed by the relatively small sample size. Additional results promote the use of specific SVM architecture (C-SVM and RBF kernel) rather than others (Nu-SVM and polynomial kernels) for performance and computational cost. Finally, a brief analysis of variable importance highlights the role of physiology, age, neurological status, admission variables and comorbidities in perfect agreement with the literature. In the following chapter an additional layer of customization will be implemented, which consists of exploring the predictive value of additional groups of variables. Doing so will slightly modify the nature of the problem by altering the variables to observation ratio significantly. The behaviour of the techniques presented here will consequently be further explored in this new context and additional tools for the estimation of variable importance will be presented.

Chapter 5

Predictive power of additional features derived from the data

Introduction

The previous chapter has revealed the importance of adequate variable pruning in the context of limited data availability. This work was carried out on a limited subset of variables that are included in the APACHE-IV model, which relies to a large extent on expert knowledge (choice of clinical variables to be collected and to be included in the final model). Nonetheless, the presentation of the MIMIC-II database in chapter 2 showed that thousands of features are actually collected at the bed side. The relative contribution of these clinical variables for the prediction of adverse outcomes in the ICU could be investigated by appropriate use of the techniques presented in chapter 4. In this chapter we therefore independently evaluate the importance of different groups of clinical features with respect to the estimation of patient severity. In the first instance, a brief introduction to *relevance* – any metrics related to the importance of a feature – will be provided, to supplement the information provided in section 4.1.3. Then, all covariates presented in the database will be used as input features to study the additional predictive information they may bring. Similarly, the potential of physiologically meaningful non-

linear combinations will then be estimated. Finally, the predictive power of dichotomous variables coding for the presence of missing values will be presented. Eventually, all of the features will be used together to select objectively those that offers the best model.

5.1 Relevance or the estimation of variable importance

We have seen in the previous chapter that the weights given to different variables in logistic regression - with or without the LASSO - can be used as an estimate of variable importance as long as all of the input variables are normalized before training. In fact, we also showed in section 3.1.1 how these coefficients can translate to an increased risk per unit change in the variable: the odds ratio. Alternative ways to derive relevance exist for other modelling techniques such as an ensemble of trees (RF and the BEF), which we introduce here.

5.1.1 Variable importance from random forest

In the RF, Out Of Bag (OOB) observations (those that were not used to build the tree) are first passed down every tree k in the forest to count the number of correctly identified samples $N(k)$. This is repeated after randomly permutating a variable m giving $N^{(m)}(k)$. The raw variable importance $\mathcal{R}^{(m)}(\mathrm{k})$ at each tree k is simply obtained by subtracting the number of correctly identified OOB observations with random permutations from the number of correctly identified OOB observations without random permutations:

$$\mathcal{R}^{(m)}(k) = N(k) - N^{(m)}(k) \tag{5.1}$$

Assuming independence of the trees [44], the standard errors (defined by $SE = \sigma^{(m)}/\sqrt{n_{tree}}$) can then be easily derived and used to compute a Z-score

$$Z = \frac{\mathcal{R}^{(m)}}{SE} \tag{5.2}$$

This Z-score is assumed to be normally distributed, which allows the derivation of a measure of statistical significance.

5.1.2 Variable importance for Bayesian ensemble of forests

The obvious drawback of this approach is that random permutation leads to different results at each attempt. To prevent this, a different approach was preferred for the BEF where the relevance criterion was defined as

$$\mathcal{R}^{(m)} = \mathcal{L}(\mathcal{F}(X)) - \mathcal{L}(\mathcal{F}(X^{(m-)})) \tag{5.3}$$

where $\mathcal{F}(X)$ denotes the output of the ensemble of forests as described in equation 4.27 and $\mathcal{L}(\mathcal{F}(X^{(m-)}))$ the likelihood of the forest with all values of X for variable m set to zero.

5.1.3 Other techniques for the estimation of variable importance

The relevance criterion of ReliefF (defined in algorithm 4.1.1) provides an additional way to estimate the relevance of input features, hence giving a total of four independent estimates from LASSO, RF, BEF, and ReliefF. Similar approaches have also been implemented for other modelling techniques, in which hyper-parameters can be added to weight each input variable: Relevance Vector Machine (RVM) for SVMs [272] and Automatic Relevance Determination (ARD) for NNs [186]. The mathematical elegance of these approaches however often comes at the price of additional assumptions (normality of input variables) and a potential decrease in performance, which was neither desired nor necessarily compatible with the presence of binary or highly skewed variables. The work on the Bayesian framework introduced by MacKay et al. [188] is however of particular interest. It will be further described and modified to suit the purpose of feature selection.

5.2 Including additional co-variates

According to Moreno and Apolone [202], the third level of customization for models estimating risk includes the search for more adequate predictors. For instance, the recent progress in the management of respiratory failure may translate to a decrease in the predictive power of this variable over time. In a context of limited data availability for model design, the replacement of this covariate by another may be indicated. Similarly, the rules for admission to the ICU have evolved over time so that the traditional variables reflecting the population's physiology may need to be updated. In particular, population ageing directly translates to an increasingly older population at admission to the ICU for whom the markers of severity are likely to be different from the markers appropriates for their younger counterparts.

For the problem described in this thesis, the need for customization is exacerbated by the fact that the population described in chapter 2 is not a general ICU population. In particular, the markers of severity are expected to be specific to a population of severe sepsis patients for whom treatment of hypotension was required. To explore this hypothesis, all of the available covariates presented in section 2.3 will be passed down to the machine learning techniques described in chapter 4. This will allow the objective identification of the optimal features for the modelling of patient severity. The novel covariates include: additional physiological markers from laboratory results and bedside monitors, some comorbidities, treatments (isotonic solutions, colloids, lactate ringers, vasopressors), microbiology results, and fluid balance (urine output). The relative importance of these different groups of variables will be discussed further.

All variables presented in section 2.3 will be used as input variables in this section: chronic health conditions (CHC), demographic data, admission variables, physiological variables, neurological status, microbiology results, and medications and interventions.

5.3 Physiologically meaningful non-linear combination of raw variables

The complexity of human physiology was touched on in sections 1.2.1 and 1.2.2. The relation between the innate and the adaptive immune systems, their relation to the inflammatory response, and the cardiovascular function are mostly directed by non-linear feedback loops. This is an essential feature of physiology that certainly should guide the development and choice of engineering techniques designed to investigate human physiology. Some of the non-linear indices related to the severity of different physiological systems are introduced in this section.

In addition to this, there is a possibility that flexible non-linear machine learning techniques such as neural networks or forests are capable of identifying non-linear relations between features; for instance, squared heart rate divided by diastolic blood pressure. While we will not challenge this view, we believe that knowledge-based strategies may ease the process in a data constrained environment. To demonstrate this, interesting non-linear combinations found in the literature were implemented and used as features presented to feature selection techniques.

5.3.1 Body indices

In this section all weight (w), height (h), and age values were given in kilograms, centimetres, and years, respectively.

Body surface area The body area surface (BSA) is a long-standing aggregated value of height h and weight w [124], which was recently redefined by Schlich et al. [250] thanks to modern imaging as

$$\text{In males} \qquad BSA = 9.75482 \cdot 10^{-4} \times w^{0.46} \times h^{1.08} \tag{5.4}$$

$$\text{In females} \qquad BSA = 5.79479 \cdot 10^{-4} \times w^{0.38} \times h^{1.24} \tag{5.5}$$

The validity of such an index has recently been challenged following a series of reports suggesting a protective effect of being slightly overweight in various conditions, including in the ICU. This value is also regularly used to compute several other indices involved in the titration of medication (drug dosage).

Ideal weight is similar to the body mass index and is defined by Robinson et al. [245] who suggested an ideal healthy weight as a function of height:

$$\text{In males} \qquad iw = 50 + 0.91 \times (h - 152.4) \tag{5.6}$$

$$\text{In females} \qquad iw = 48.67 + 0.89 \times (h - 152.4) \tag{5.7}$$

this formula is also used for the titration of drugs and in particular in sepsis [247]. The feature we included in our study was $\delta w = w - iw$.

5.3.2 Kidney indices

BUN to creatinine ratio Urea nitrogen is a bypass product of the digestion of protein that is produced at a constant rate by the liver. Similarly, creatinine is a bypass product of muscle metabolism that is also produced at a regular pace. Both Blood Urea Nitrogen (BUN) and creatinine are filtered out by the kidneys and elevated blood levels therefore lead to suspected renal failure. The re-absorption mechanism however differs, which is why the ratio of the two is found to be indicative of early acute kidney injury.

Urine derivative Urine derivative $\partial U/\partial t$ is simply the temporal derivative of the instantaneous urine output signal. A positive value indicates an increased urine output from the previous recording which may serve as a proxy for kidney function. Conversely, a negative reading may indicate a degrading kidney function.

Creatinine clearance Indirect markers of renal function such as BUN and creatinine may not increase until the presence of a large disturbance of kidney dysfunction. Indi-

rect markers such as the rate at which creatinine is cleared from the blood (creatinine clearance) may ease the early identification of failure. Ideally, this can be measured directly from the creatinine levels found in the urine, but estimations exist when this is not available. In particular, Cockcroft and Gault [65] suggested the following equations :

$$\text{In males} \qquad C_{cr} = 1.23 \cdot w \frac{140 - a}{88.5 \cdot c} \tag{5.8}$$

$$\text{In females} \qquad C_{cr} = 1.03 \cdot w \cdot \frac{140 - a}{88.5 \cdot c} \tag{5.9}$$

$$\tag{5.10}$$

for $18 < a < 110$, $25 < w < 120$, $0.6 < C \leq 7$, and where a and C denote age and creatinine blood level, respectively. Creatinine (Cr) and sodium (Na) are expressed in milliequivalents per liters (mEq/l), a unit of electric charge volumetric density commonly used for biological results and that relates to the concentration of the product tested.

Hypernatremia is characterized by an excess of sodium in the blood, which is often caused by a free water deficit in the body. Such a condition may occur in hypovolaemic sepsis patients resuscitated with saline solutions [134]. An estimation of its severity was given by Adrogué and Madias [3]:

$$H_2O_{deficit} = d \cdot w \cdot \left(\frac{Na}{140} - 1 \right) \tag{5.11}$$

where $d = 0.6$ and 0.5 in males and females, respectively.

5.3.3 Respiratory indices

PaO_2 to FiO_2 ratio is the ratio between the inspired oxygen levels (FiO_2) and the partial pressure of oxygen in the arteries (PaO_2) as a proxy for the lung capacity to transfer oxygen from the air to the lungs. Because of the ease of use, this ratio has long been exploited as an indicator of acute respiratory distress.

Alveolar-arterial Oxygen gradient (DA-aO_2) is measured to help clinicians identify the cause of hypoxaemia[1]. Estimation of the oxygen gradient is provided by Mellemgaard [195] with the following equations, that were customized the altitute of BIDMC at Cambridge, Massachusetts ($P_{atm} = 30$mmHg) :

$$\begin{aligned} DA - aO_2 &= P_AO_2 - P_aO_2 \\ &= FiO_2(P_{atm} - P_{H_2O}) - \frac{P_aCO_2}{0.8} \\ &= FiO_2 \times 713 - 1.25 \times PaCO_2 - PaO_2 \end{aligned} \tag{5.12}$$

where P_AO_2 and P_aO_2 denote the airway and alveolar pressures expressed in millimetres of mercury (mmHg), respectively. FiO_2 is a measure of the fraction (%) of oxygen present in the air inspired by the patient.

5.3.4 Cardiovascular indices

Cardiac output (CO) can be estimated to a constant as the product of the heart rate (HR) and the pulse pressure (PP, defined in section 2.3) $CO = PP \cdot HR$. This relation has been validated in healthy volunteers but it remains unclear how much the approximation holds in unhealthy people.

Cardiac index (CI) is a proxy for the volumetric blood flow; the flow (in L.min^{-1}) with respect to the volume of tissue to be perfused. It is expressed as $CI = CO/BS$, with BS the body surface defined above and where cardiac output can be the measured or derived from other observations as seen above.

Double product (DP) is the product of heart rate and systolic blood pressure ABP_{Sys} defined as $DP = HR \times ABP_{Sys}$. Interestingly, it was noted in various studies as an indicator of cardiovascular reserve [252] and death [234].

[1]Hypoxaemia indicates a low oxygen levels in the blood. It has to be distinguished from hypoxia that indicates a localized inadequacy between metabolic needs and supply.

Shock index (SI) Similarly, the shock index $SI = HR/ABP_{Sys}$ is a non-linear aggregate of systolic blood pressure and heart rate which was found to reflect the severity of shock in numerous studies [233, 243].

Stroke volume (SV) represents the volume ejected from the left ventricle at each heart beat. Clearly there is no way to access this value directly and it is commonly estimated by $SV = CO/HR$.

5.3.5 Fluid balance and treatments

Unlike most physiological values, treatments and fluids need to be integrated over a time window to be meaningful across a population. A simple way of doing this is to take the sum of the treatment over the period considered. Because treatments are already available in normalized values per weight unit, there is no need to account for variation in a patient's size. Likewise for fluids, the cumulative volume of different types of fluids will be used as an input variable.

5.3.6 Severity scores

Clinical severity scores are combinations of input variables and could therefore be considered as independent features. The Van Walraven score for instance aggregates dozens of dichotomous comorbidities within one CHC index thanks to the information derived from one hundred thousand previous cases [295]. As seen in section 3.2.1, the resulting score offers a higher performance ($Nll = 417.7 \pm 5.8$, $AUC = 62.6 \pm 3.1$) than a model using individual CHC and deriving weights for the dataset described in this work. This indicates that coefficients defining the projection of all comorbidities into a single-dimensional decision space contain interesting discriminatory information. Ideally, we would like to re-derive these coefficients for our population. Unfortunately, in a context of limited training data availability, the use of these external coefficients may not prove

useful. We will therefore include these scores (APS, SAPS, APACHE, SOFA, and Van Walraven) in the analysis and see the extent to which they contribute to model performance.

5.4 Values not missing at random

Finally, medical data is specific in the sense that the presence (and thereby absence) of a measurement can reflect the clinician's decision to order (or not) a test. This is particularly true in the ICU where data is said to be "*not* missing at random" [263]. The absence of a value may therefore be independently related to severity. Imputing a missing value with the mean of the variable only gives half of the story: this measurement is non-informative or biased towards normality. A good strategy to account for the non-randomness of missing values consists of adding a novel binary covariate indicating whenever the value is missing. This strategy potentially doubles the amount of covariates, which can be handled by a feature selection technique.

5.5 Predictive power of the novel groups of variables

5.5.1 Materials and methods

For this study, all of the data will be considered up to 24 hours after the admission time. There are two reasons to restrict our analysis to the first 24 hours of ICU stay. First of all, a fair comparison to traditional severity scores can only be achieved with an equivalent time window. Secondly, the reason why traditional severity scores actually constrain their analysis to the first day is because they are designed for benchmarking. Consequently these scores try to capture solely the patient's severity while later information is assumed to reflect treatment and intervention. This chapter will report on whether it is reasonable to assume that treatments, interventions, and their incidence can be ignored during the first 24 hours of ICU stay.

All variables observed more than once will be aggregated with the following func-

tions: minimum, mean, maximum, and standard deviation, resulting in 4 features per variable. The subsets of features considered will then be:

- raw: all raw variables (281 variables);
- $nlin$: raw and their non-linear transformations (336 variables);
- $score$: $nlin$ + Van Walraven, APS, and SOFA (339 variables);
- $missv$: missing values in raw (74 variables);
- all: $nlin$ + $missv$ (414 variables);

Exploring the potential predictive power of these large groups of novel features is complicated by the relatively small sample size of the dataset. In fact, it only is possible to do so with the help of appropriate machine learning techniques such as that described in chapter 4. The evaluation of the subset of variables included in the APACHE-IV model (chapter 4) revealed that simpler models rank best. This can be explained by the fact that fewer parameters can be more accurately be determined in a data-constrained environment. In particular, LASSO seems to offer an ideal trade-off between simplicity, computational cost, and performance. Additionally, the final model it provides (logistic regression model) is the most broadly applied technique for the estimation of risk in the ICU. Consequently, we will use the LASSO to explore the contribution of different groups of variables without too great a computational cost.

Yet, adding a large amount of covariates to the problem somehow changes its nature and shifts the focus from pruning to proper feature selection techniques. In other words, the problem becomes more about identifying a small subset of features within many rather than discarding a few within a few. The same theory addresses these problems (feature selection techniques) but it is reasonable to assume that different techniques will behave differently in these two contexts. Consequently, the experiment which includes all the available covariates will be ran with the techniques presented in the previous chapter, except for the cascaded-SVM [62] and Nu-SVM that did not compare advantageously to C-SVMs.

Table 5.1: *Performance of models developed using variables collected during the first 24 hours of ICU admission expressed with the Negative Log-Likelihood (Nll), the AUROC, and the Standardized Mortality Ratio (SMR). Different models are considered with different subset of variables: all raw variables (raw), raw and their non-linear transformations $nlin$; raw converted into binary variables indicating where values are missing ($missv$); and $nlin$ in addition to APS, SOFA and Wan Walraven ($score$). The top rows shows the performance of different models considering all variables and the bottom rows shows the performance of LASSO using using different groups of variables.*

Features	Models	Type	Nll	AUROC	SMR
all ($k = 414$)	LASSO		337.8±6.7	80.39±0.27	0.9±0.1
–	BEF		338.6±8.9	81.01±1.50	1.0±0.0
–	SVM	Nu-SVM (RBF)	341.6±10.7	79.98±1.85	1.0±0.1
–	SVM	Nu-SVM (linear)	344.9±10.7	80.11±1.43	0.9±0.1
–	SVM	C-SVM (RBF)	346.2±10.7	79.79±1.38	0.9±0.0
–	SVM	C-SVM (linear)	351.2±6.0	79.08±0.23	0.9±0.1
–	SVM	C-SVM (polynomial)	377.1±7.2	76.98±0.80	1.0±0.0
–	RF	$m = M/10$	380.4±5.5	77.76±1.23	0.9±0.1
–	SVM	Nu-SVM (polynomial)	382.7±10.2	77.17±0.47	0.9±0.1
all ($k = 414$)	LASSO		337.8±6.7	80.39±0.27	0.9±0.1
$score$ ($k = 339$)	–		339.7±8.7	80.20±0.88	0.9±0.0
$nlin$ ($k = 336$)	–		346.3±8.5	79.23±0.55	1.0±0.0
raw ($k = 281$)	–		348.7±10.1	78.85±0.83	0.9±0.0
$missv$ ($k = 74$)	–		386.1±3.9	71.38±0.94	0.8±0.1

Abbreviations: Intensive Care Unit (ICU), Acute Physiology Score (APS), Sepsis-related Organ Failure Assessment (SOFA), Least Absolute Shrinkage And Selection Operator (LASSO), Negative Log-Likelihood (Nll), Area Under the Curve (AUROC), Standardized Mortality Ratio (SMR), Bayesian Ensemble of Forests (BEF), Random Forest (RF), Radial Basis Function (RBF)

All covariates' weights will be estimated within a three-fold validation procedure using the same split as in the previous chapters. The estimated probabilities will be generated for each model (at each fold) and then used to compute metrics of performance, which will be characterized by their mean and standard deviation. Finally, each model will be applied to the entire dataset in order to identify parameters and the importance of feature for all available data.

5.5.2 Results

Table 5.1 presents the results of LASSO using different subsets of features (bottom) and all models with all variables (top). The table provides valuable information on the use-

fulness of features and techniques for the problem we describe.

First and most importantly, additional covariates clearly improve the prediction of mortality from $Nll = 369.4$ and $AUC = 75.1\%$ with APACHE-IV covariates to $Nll = 348.7$ and $AUC = 78.9\%$ with all raw covariates extracted from MIMIC-II using the LASSO model. Then, there seems to be an interest in using the additional groups of features suggested even though the difference between the different subgroups (raw, $nlin$, $score$, and all) was not found to be statistically significant. Non-linear combination of these raw variables offers an additional gain in discriminative power ($IDI = 0.8, p = 0.13$) from which a similar improvement is provided by the addition of severity scores as an input feature ($AUC = 80.2\%$). Finally the use of all input features (except severity scores) offered the best performance with $Nll = 337.8$ and $AUC = 80.39\%$ providing an integrated discriminatory improvement of $IDI = 3.3, p = 0.02$ when compared to the best results reported with the APACHE-IV subset of variables in the previous chapter.

Interestingly, a model using only binary values to indicate whether a value is missing or not offers an AUC of 71.38%. This indicates that values are not missing at random. This phenomenon has long been described [209] and derives from a simple fact: every intervention has a cost (financial or medical) which also applies to tests. For instance, Magnetic Resonance Imaging (MRI) is an expensive and limited resource that clinicians restrict to patients who are mostly likely to benefit from it. Similarly, blood samples are minimized so as to not further reduce a patient's volume, which particularly applies to patients with sepsis who already lack fluids. In both cases, the presence (or not) of a variable directly reflects the clinician's opinion of the patient's physiology.

Again, of all the machine learning techniques tested on this dataset (with $k = 414$) LASSO performed the best with the BEF. No statistically significant difference was found between the different SVM architectures although the polynomial kernel consistently performed worse. Random forest classifiers provide better results on this dataset than in previous chapters but does not compare favourably with other techniques including BEF. This corroborates the recommendation of Breiman [44] who suggested using a feature

importance metric to filter out meaningless variables prior to growing the forest.

5.5.3 Most predictive features

Table 5.2 shows the thirty most important predictors identified from the different groups of variables (raw, $missv$, $nlin$, $score$, and all) and according to the different modelling techniques (LASSO, BEF, RF, and ReliefF). Figure 5.1 offers a more accessible although less accurate presentation of the data presented in the last four columns of table 5.2. First of all, the left-hand column shows that additional clinical features complement the groups of variables traditionally included in severity scores such as SAPS-3 and APACHE-IV. For instance, the average urine output recorded, the presence of chronic depression, the level of alkaline phosphatase (ALP), the blood levels of chlorides, and ethnicity are usually not included in severity scores. In addition to this, the most relevant variables seem especially specific to the population of patients with sepsis and hypotension. For instance, temperature, heart rate, respiratory rate, and WBC all relate to the SIRS. Lactate is a by-product of anaerobic metabolism and therefore indicates a lack of tissue perfusion. SOFA components and urine output indicate some extent of organ dysfunction. Finally, the time of onset and the length of the hypotensive episode are specific to the sepsis-induced hypotension. The inclusion of all these variables in addition to the increase in classifier performance of the model built from these covariates clearly demonstrates the importance of this level of customization.

Table 5.2: *Thirty most important variables according to relevance metrics derived from different models and ensembles of variables. The models are: Least Absolute Shrinkage And Selection Operator (LASSO), Bayesian Ensemble of Forests (BEF), Bayesian Ensemble of Forests (BEF), and ReliefF. The groups of variables are: all raw variables (raw), raw and their non-linear transforms $nlin$; raw converted into binary variables indicating where values are missing ($missv$); and $nlin$ in addition to APS, SOFA and Wan Walraven ($score$). When appropriate, the variables are aggregated over the first 24 hours with an operator: minimum (Min), Mean (μ), Maximum (Max), and standard deviation (σ). Missing values are binary coded and indicated by the letters "MV". In parenthesis is indicated the metric of relevance: the coefficient β for LASSO, $\mathcal{R}(k)$ for RF and BEF (see section 4.1.3), and the relevance index for the ReliefF algorithm (as described in algorithm 4.1.1).*

LASSO					BEF	RF	ReliefF
raw	$missv$	$nlin$	$score$	all	–	–	–
SAPS-I (0.40)	No Hepatic failure (-0.37)	SAPS-I (0.42)	SOFA Admission (0.37)	SAPS-I (0.40)	SAPS-I (429.70)	Urine Output (0.19)	Tropo-T – MV (5.39)
Urine Output – μ (-0.26)	No Bypass (0.20)	Urine Output (-0.24)	SAPS-I (0.34)	No Hepatic failure (-0.26)	No Hepatic failure (267.70)	HCO_3 – μ (0.18)	No Hepatic failure (5.29)
Temperature – μ (-0.24)	Lact. Ringers – MV (0.19)	Temperature – μ (-0.19)	Temperature – μ (-0.23)	Age (0.22)	Age (238.42)	INR – Max (0.18)	Age (5.20)
Elective Adm. (-0.19)	Cardio SOFA – MV (-0.15)	Depression (-0.17)	Urine Output (-0.21)	Urine Output (-0.18)	Urine Output - μ (196.14)	Hemato. SOFA – μ (0.17)	Hep. SOFA – μ (5.10)
Depression (-0.18)	Ca^{2+} – MV (-0.15)	Hemato. SOFA – μ (0.16)	Van Walraven (0.20)	Lact. Ringers – MV (0.18)	Onset time (145.11)	pH – Min (0.17)	SAPS-I (5.00)
HR – Max (0.16)	Not Intubated (-0.14)	Met. Cancer (0.16)	Lactate – Min (0.17)	Hemato. SOFA – μ (0.16)	Temperature – Max (140.99)	SAPS-I (0.17)	SaO_2 MV (4.81)
Lactate – Min (0.16)	Age – MV(0.11)	Elective Adm. (-0.15)	APACHE-IV (0.17)	Met. Cancer (0.16)	Urine Output (125.69)	Urine Output - μ (0.16)	CHF (4.69)
Met. Cancer (0.16)	$PaCO_2$ – MV (-0.10)	Lactate – Min (0.15)	Lact. Ringers – $\sum$ (-0.16)	Temperature – μ (-0.16)	Temperature – μ (114.50)	No Hepatic failure (0.16)	Liver Disease (4.69)
Hemato. SOFA – μ (0.15)	Tropo-I – MV (-0.10)	Age (0.14)	Depression (-0.15)	Depression (-0.14)	Met. Cancer (67.49)	Bilirubin – Min (0.16)	CVP – MV (4.68)
Age (0.15)	INR – MV (-0.09)	HR – Max (0.14)	Chronic Pulm. (-0.13)	Lactate – Min (0.13)	Hemato. SOFA – μ (63.94)	BUN – Max (0.16)	Is Not Intub. (4.47)
ALP – Min (0.15)	Cholesterol – MV (0.09)	Hepatic SOFA – μ (0.13)	Elective Adm. (-0.13)	No Bypass (0.13)	Insulin – $\sum$ (58.42)	Bilirubin – Max (0.16)	Not Staph. Aur. Coag. (4.36)
Surgical ICU (-0.14)	WBC – MV (0.08)	Lact. Ringers – $\sum$ (-0.13)	Insulin – $\sum$ (-0.12)	Elective Adm. (-0.12)	Lactate – Min (54.42)	BUN – Min (0.15)	Eye Open (4.24)
BUN – Min (0.14)	SAS-I – MV (0.08)	Insulin – $\sum$ (-0.13)	Surgical ICU (-0.11)	HR – Max (0.11)	GCS – μ (53.47)	Hepatic SOFA – μ (0.15)	Hema. SOFA (4.12)
No Bypass (0.14)	No Insulin (0.07)	ALP – μ (0.12)	ALP – Min (0.11)	BUN – Min (0.10)	HR – Max (53.33)	Urine – Max (0.15)	Not Blood Infection (4.00)
Hepatic SOFA – μ (0.14)	No Sedative (0.07)	Shock Index - Max (0.12)	FiO_2 – σ (-0.11)	Ca^{2+} – MV (-0.10)	Ethnicity (53.09)	Hepatic Failure (0.15)	Surgical ICU(3.87)
Resp. Rate – Min (0.12)	SaO_2 – MV(-0.06)	Ethnicity (0.10)	Age (0.11)	Ethnicity (0.10)	BUN/Creat – μ (52.31)	Bilirubin – μ (0.14)	Albumin – MV (3.74)
INR – μ (0.12)	Tropo-T – MV (-0.06)	INR – μ (0.10)	WBC – σ (-0.10)	$PaCO_2$ – MV (-0.10)	Lact. Ringers – MV (51.09)	Lactate – Max (0.14)	Trop-I – MV (3.61)
Bilirubin – Max (0.11)	Bilirubin – MV (-0.06)	Bilirubin – Max (0.10)	Hemato. SOFA – μ (0.10)	Age – MV(0.10)	Lact. Ringers – $\sum$ (50.30)	AST – μ (0.13)	Weight Loss (3.46)
Liver Disease (0.11)	Resp. SOFA – MV (-0.06)	Resp. Rate – Min (0.10)	Ethnicity (0.10)	INR – μ (0.09)	Eye Open (49.41)	Lactate – Min (0.13)	Renal failure (3.32)
Motor Response (-0.11)	No Enteroco. Faecium. (-0.06)	BUN – Min (0.10)	Resp. Rate – Min (0.10)	Liver Disease (0.09)	Resp. Rate – Min (47.55)	INR – μ (0.13)	No Sedative (3.16)
Chloride – Max (-0.11)	ALT – MV (-0.05)	Liver Disease (0.09)	INR – μ (0.09)	Bilirubin – Max (0.08)	Resp. Rate – μ (46.59)	Bilirubin – σ (0.13)	Met. Cancer (3.00)
FiO_2 – σ (-0.10)	No Graft (0.05)	FiO_2 – σ (-0.09)	Shock Index – Min (0.09)	DP - μ (-0.08)	HR – μ (44.80)	Verbal Resp. (0.13)	FiO_2 – Max (2.98)
Ethnicity (0.10)	Not Staph. Aur. Coag. (-0.05)	Surgical ICU (-0.09)	BUN/Creat – Min (0.09)	pH - σ (0.08)	pH - σ (42.34)	Eye Open (0.13)	Motor Response (2.65)
Pressors – σ (0.10)	FiO_2 – MV (-0.05)	BUN/Creat – Min (0.09)	Liver Disease (0.09)	Eye Open (-0.08)	INR – μ (42.19)	Platelets – Min (0.13)	Eye Open (2.60)
WBC – σ (-0.10)	Not Pseudo. Aerus (-0.05)	Resp. Rate – μ (0.09)	Eye Open (-0.09)	BUN/Creat – Min (0.08)	BUN/Creat – Min (41.71)	Urine – σ (0.13)	Urine – σ (2.50)
HR – μ (0.09)	Glucose – MV (-0.05)	Chronic Pulm. (-0.08)	Bilirubin – Max (0.08)	WBC – σ (-0.08)	WBC – σ (41.08)	ALT – Max (0.13)	ALT – Max (2.47)
Chronic Pulm. (-0.09)	Lactate – MV (-0.04)	WBC – σ (-0.08)	Temperature – σ (-0.08)	No Insulin (0.07)	BUN – Min (39.35)	Temperature – μ (0.13)	Temperature – μ (2.45)
Temperature – σ (-0.09)	No Riker's Scale (0.04)	Pressors – σ (0.08)	Met. Cancer (0.08)	Hepatic SOFA – μ (0.07)	Pre-LOS (39.27)	pH – MV (0.12)	pH – MV (2.02)
Eye Open (-0.09)	HCT – MV (-0.04)	Eye Open (-0.08)	Na – Min (0.08)	Chronic Pulm. (-0.07)	Temperature – σ (39.23)	Lactate – μ (0.12)	Lactate – μ (2.02)
Mg – Max (-0.08)	NIMAP – MV (-0.04)	Chloride – Max (-0.08)	Hepatic SOFA – μ (0.08)	Shock Index – Min (0.07)	SpO_2 – σ (36.24)	pH - σ (0.12)	pH - σ (2.01)

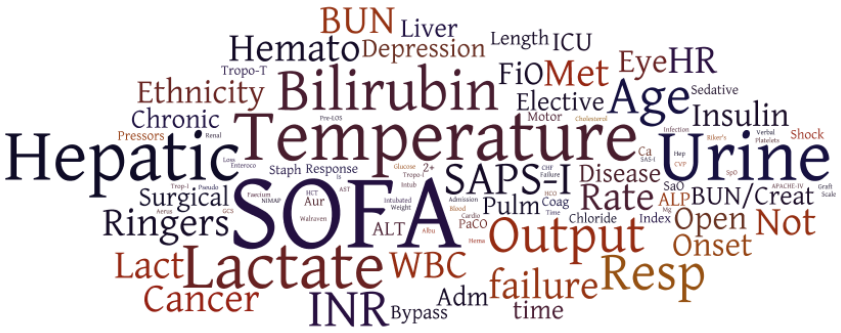


Figure 5.1: *Aggregated representation of variable importance. The size of each variable name on this "word cloud" is the mean of the normalized variable importance metric given by the LASSO, BEF, RF, and ReliefF algorithms as seen on table 5.2.*

The coding of missing values into binary variables also provides interesting results. For instance, it is known that hypocalcaemia is linked to very poor outcomes in critically-ill patients [82] and in particular in patients with sepsis [314]. Of all electrolytes, only the presence of a Ca^{2+} reading during the first 24 hours (reflecting the suspicion of Ca^{2+} derangement) was found to be associated with severity (positively). Similarly, the presence of at least one Arterial Blood Gas (ABG) reading such as $PaCO_2$ or SaO_2 was also found to relate positively to severity. Finally, results from microbiology indicating the presence of some resistant agents, Enterococus Faecium [108] and the well known meticillin-resistant Staphylococcus aureus (MRSA) [275], were also found to increase severity.

In addition to the importance of missing values found in blood tests, missing values also reveal the importance of interventions (prior and during the ICU stay) in the prognosis of sepsis. The presence of by-pass surgery and any number of grafts prior to admission were associated with decreased severity, which certainly relates to the less severe group of post cardiac surgery patients. Similarly, the presence of some interventions during the first 24 hours was also found to relate to severity: lactate ringers fluids, intubation, insulin, and sedatives; of these, only intubation was found to be positively correlated with severity. The role of insulin in the treatment of sepsis is somewhat controversial: an in-vitro study demonstrated sepsis-induced insulin-resistance in a murine model [121] while the use of intensive insulin therapy was later discouraged in a population of patients with severe sepsis [46]. These findings however stress the importance

of intervention variables during the first 24 hours of ICU admission and challenge the hypothesis that the data collected in this time window can be used to build models for benchmarking purposes.

The contribution of a non-linear combination of variables was also found to be beneficial. First, the integration (sum of) of treatments over the period considered contributes to the model as with lactate ringers and insulin. Only two other non-linear indices contributed significantly ($p < 0.05$ in the LASSO model): BUN to creatinine ratio and shock index. In terms of severity scores, APS and SOFA contributed equally to the model and should be included in the future analysis (they are not part of the group all).

Finally, the identification of features of interest from the most complete subgroup (all) with different modelling techniques revealed some interesting commonalities and variations. First of all there is a consistent group of variables which came out top: SAPS-I, age, hepatic failure, urine output, and temperature. Bilirubin blood level and respiratory indices also play an important role, which is consistent with findings reported in the literature, as presented in section 3.2.1. Metastatic cancer and hepatic failure were the only two comorbidities consistently selected by different techniques. To summarize, a data-driven approach, which is not dependent on prior known biology, enabled us to select the best predictors for inclusion in our model.

5.6 Conclusion

The previous chapters suggested that the customization of a model's parameters to a specific population improves the overall performance when using the clinical covariates included in APACHE-IV. In this chapter we explored the potential of a second level of customization: the addition of novel covariates to build the model. First, all available clinical observations were used to derive the model. Then, different subgroups of variables were added: non-linear combinations of clinical covariates (including existing severity scores) and missing values. To deal with such large sets of variables and identify important

clinical predictors in each subgroup, different estimations of variable relevance were introduced for each machine learning technique implemented. Again, the simpler model (LASSO) compared advantageously to more complex techniques (RF, BEF, SVM) that did not improve the estimation of a patient's severity. Taken independently or together, each group of covariates contributed positively to the model highlighting the importance of some variables with potentially meaningful clinical implications.

Chapter 6

Feature selection and parameter pruning using a genetic algorithm

6.1 Introduction

In the preceding chapters we have seen that the problem of outcome prediction in the ICU is complicated by the large number of available covariates and the limited number of observations in existing databases. Nonetheless previous results have also built a strong rationale for the use of objective variable selection rather than one based purely on expert knowledge. The machine learning community has addressed the problem of feature selection with encouraging results despite a slight predominance of filter techniques such as ReliefF and mRMR that seem less optimal and elegant than wrapper techniques such as LASSO. In addition to this, there is much evidence that grid searches for the identification of model hyperparameters are not an optimal solution, from the point of view of both performance and computational cost [28]. To illustrate this, the use of suboptimal pruning technique (grid-search) and filter feature selection technique (ReliefF) for SVMs optimization possibly account for the relatively low performance reported in chapter 4 (AUC of 74.3%). Consequently, a global framework for parameter pruning and feature selection may improve the performance of these techniques.

There is a broad variety of heuristic algorithms for optimization, which could potentially serve as a global framework for this task. In particular, nature offers numerous examples of how the optimization of global functions using distributed and local intelligence: ant colonies for the search and processing of food sources [79]; various swarm organizations such as fishes, bees, and birds [149]; and social spiders that use a series of local rules to build massive webs [71]. From these, a whole variety of bio-inspired techniques have been created, most of which are massively distributed. For instance, Particle Swarm Optimization (PSO) [154] consists of a swarm composed of particles representing a given combination of parameters in the search space. After random initialization, particles are iteratively moved through the search space at random and according to local (position and direction of neighbours) and global rules (centre of mass of the swarm). These rules are inspired by swarm movements in nature and are defined with respect to an objective or *fitness* function that has to be optimized. Eventually, the process converges to a global solution.

A Genetic Algorithm (GA) is a heuristic optimization algorithm that mimics the mechanisms of DNA duplication and natural selection. It is initialized with random 'individuals' (also called genotypes constituting a binary vector which defines which subset of variables will be used). The performance of each individual is estimated with a fitness function (measuring how well an individual, or given subset of variables, discriminates and calibrates in a prediction task). The initial (random) choice of individuals is likely to be sub-optimal, and an iterative process of 'natural selection' is used to converge to a stable population of suitable individuals. At each iteration, a percentage of the best individuals bred to generate offspring (the new generation), to whom selection and breeding is subsequently applied. The best individuals are cloned (to ensure that the best individual in the next generation is at least as good as in the previous one), and the poorest performing individuals are removed from the "breeding cycle". Eventually, this evolutionary process selects the most adapted subset of variables with respect to the given fitness function [106, 137, 205]. This technique has been successfully applied to variable

selection [308] and in particular to biomedical and clinical datasets [315, 144].

This chapter describes the concept of GAs in more detail and presents their implementation in this work. In particular, different fitness functions are introduced to tackle the problem of combined feature selection and parameter pruning introduced above. All experiments presented in this chapter use exactly the same three-fold cross-validation procedures as in the previous chapters. The data have also been normalized as described by equation 4.32 with special care given at every cross-validation level not to input information derived from the validation set.

6.2 Description of Genetic Algorithm

6.2.1 A brief introduction to Genetic Algorithms

The terminology of GAs, like their mechanism, is largely inspired by biology and will be described here to ease the understanding of the following sections. In particular, table 6.1 lists the parameters, indices and notations that will be used throughout this chapter.

Table 6.1: *List of parameters used to describe principles of Genetic Algorithms in this section.*

Parameter	Index	Maximum value	Notation
Patient	i	N	
Variable	j	P	X,x_{ij}^b
Bootstrap	b	B	
Generation	k	S	
Individual	l	R	Π,π_{lm}^k
Gene	m	Q	

Individual In the context of a binary GA like the one described in this work, genotypes are binary vectors and will be denoted by π_l. They are also referred to as *individuals* since

different bits (or groups of bits, i.e. *genes*) code for a specific parameter. For instance, as illustrated in Figure 6.1, a binary gene π_{lm}, of length one, can code whether or not a certain variable is included in the model described by the genotype. Alternatively, another gene of a longer length can code a parameter's value after conversion from binary to decimal values.

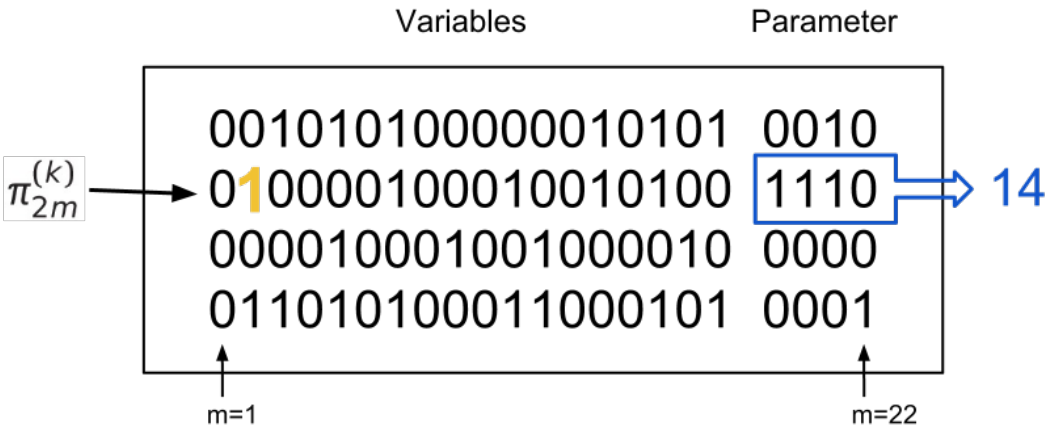


Figure 6.1: *This shows an example of a population with $R = 4$ genotypes of length $Q = 22$. The genes located $m = 1$ to 18 code for variable selection so that genotype π_{2m} selects variable 2 with location $m = 2$ set to $\pi_{22} = 1$ (yellow). The gene located from $m = 19$ to 22 codes for a parameter that is then converted to decimal: in blue $\pi_{2[19..22]} \equiv 14$.*

Population The genotypes constitute the lines of a matrix, denoted Π. This constitutes the "population" as it is composed of several individuals. Further in this section, we will see that GAs are iterative algorithms and a population will consequently also be referred to as a *generation*. At each generation, the best individuals with respect to a *fitness function* will be paired up and bred to create a new generation (i.e. the offspring). Over generations, the performance of the fitness function will be increased to finally select genotypes (combination of variables and parameters) that offer the best performance.

6.2.2 Initialization

The first population, $\Pi^{(0)}$, is a random population of R genotypes. In this work, genes coding for variable selection are initialized with a Bernoulli distribution with the mean ar-

bitrarily set to $p = N_{training} * P(Non-surviving)/10)$ where $N_{training}$ denotes the number of observations in the training set and $P(Non-surviving)$ the probability of a patient not surviving on this set. This initialization ensures that, on average, genotypes select about one variable per 10 positive occurrences of the event (in-hospital mortality), thereby ensuring a favourable number of cases to variables ratio. On the other hand, genes coding for parameters are drawn from a Bernoulli distribution with the mean arbitrarily chosen as $p = 0.3$; this parameter however was not found to influence the algorithm behaviour.

6.2.3 Iteration

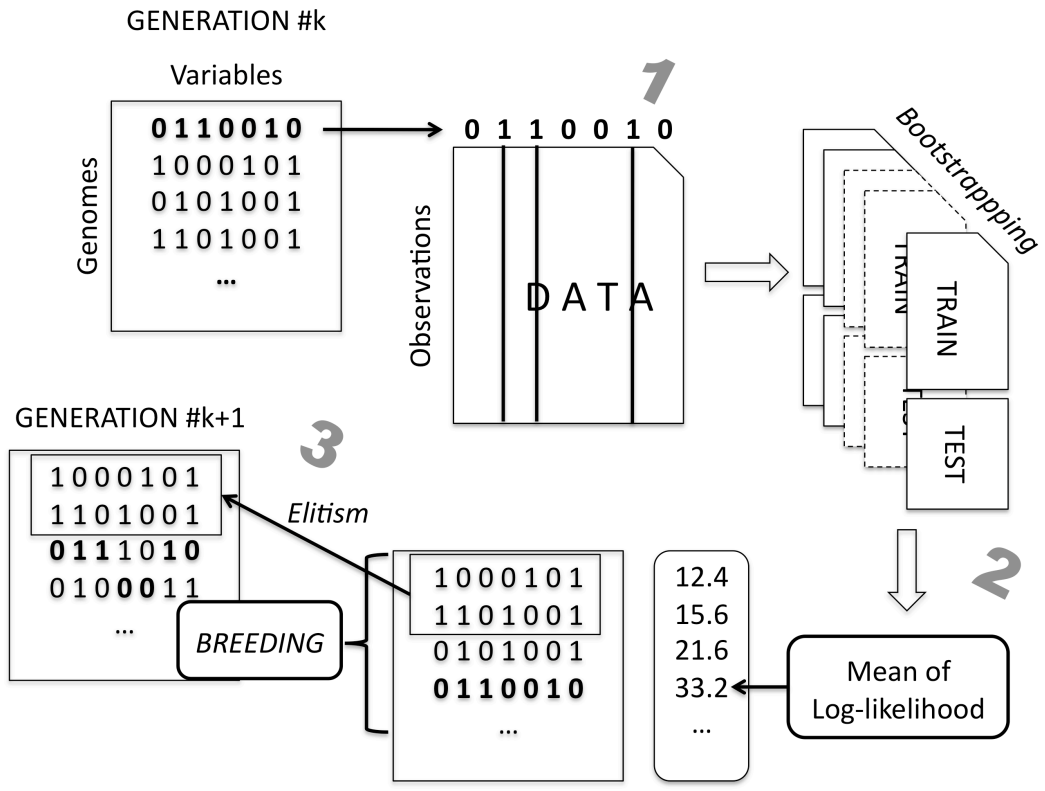


Figure 6.2: *Description of iteration k of the Genetic Algorithm. Each row (genotype/individual) is evaluated as follows: (1) features indicated by the binary genotype are extracted from the data set; (2) during a bootstrapping procedure the performance of the subset of variables is estimated on different validation sets; finally (3) the new generation $\Pi^{(k+1)}$ is created as detailed in figure 6.3.*

Each generation $\Pi^{(k)}$, starting from initialization, is then evaluated as described in

figure 6.2. For each genotype:

1. The subset of features indicated by each genotype $\pi_l^{(k)}$ is extracted from the dataset, which incidentally reduces the dimensionality;

2. Model parameters are derived for each genotype, l, with a specific mapping function Φ that covers a desired range for the parameter; for example, if the parameter belongs to $[0, 1]$ it will be coded as: $p = \Phi(\pi_{lm}) = \frac{bin2dec(\pi_{lm})}{bin2dec(111...11)}$ where $bin2dec$ denotes the function converting a binary sequence into a real number so that for instance, $bin2dec(101) = 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 = 9$), m indicates the parameter's indices (location in genotypes) and the denominator represents the highest possible value coded;

3. Each model is then trained with the selected variables and parameters during a bootstrapping procedure, as introduced in section 3.1.4.2. The output of each model trained is then used to evaluate a fitness function that reveals how well the combination of variables and parameters is performing on "unseen" data as detailed in 6.2.3.1;

4. Finally, the population is ordered with respect to the fitness function and bred as explained in section 6.2.3.2.

6.2.3.1 Fitness function

Most machine learning techniques have to be controlled to avoid overfitting (see sections 3.1.4.2 and 4.1) and a general rule is that the more complex the model, the more likely the overfitting is to occur. To prevent the optimization process from iteratively selecting more and more complex solutions (by increasing the number of selected variables for instance) and thus overfitting the training data, two strategies can be implemented: (1) an additional cross-validation layer can be added to optimize the out-of-sample performance or (2) a penalty term can be added to the fitness function. The first solution comes with several limitations that will be further detailed while the second brings us

back to the choice of an arbitrary hyperparameter: the weight of the "shrinkage" coefficient. To conclude, depending on its nature, the fitness function can be estimated on the entire training set or on the validation set should any additional cross-validation layer be implemented at this level.

If so, for each training set b, the performance of the subset of variables is estimated as follows:

1. Model parameters $\beta^{(b)}$ are fitted to the training data;
2. The probability of a patient dying $\hat{y}^{(b)}$ is then estimated for all the cases in the validation set;
3. Finally, the performance of the genotype on the validation set $X^{(b)}_{validation}$ can be taken as the log-likelihood (equation 3.4) even though any other metric of performance can possibly be considered.

6.2.3.2 Breeding

The breeding process is detailed in figure 6.3. It shows that a child population is composed of:

- 10% of the best genotypes from the parent population (3b);
- 90% mutated offspring (3a), which were created during a three-step process, that mimics DNA replication:

3a-i genotypes falling into the best 45% of the parent population are randomly paired up;

3a-ii selected pairs of parents are crossed over at random locations, creating two children;

3a-iii finally, children undergo random mutation of 20% of their genotypes.

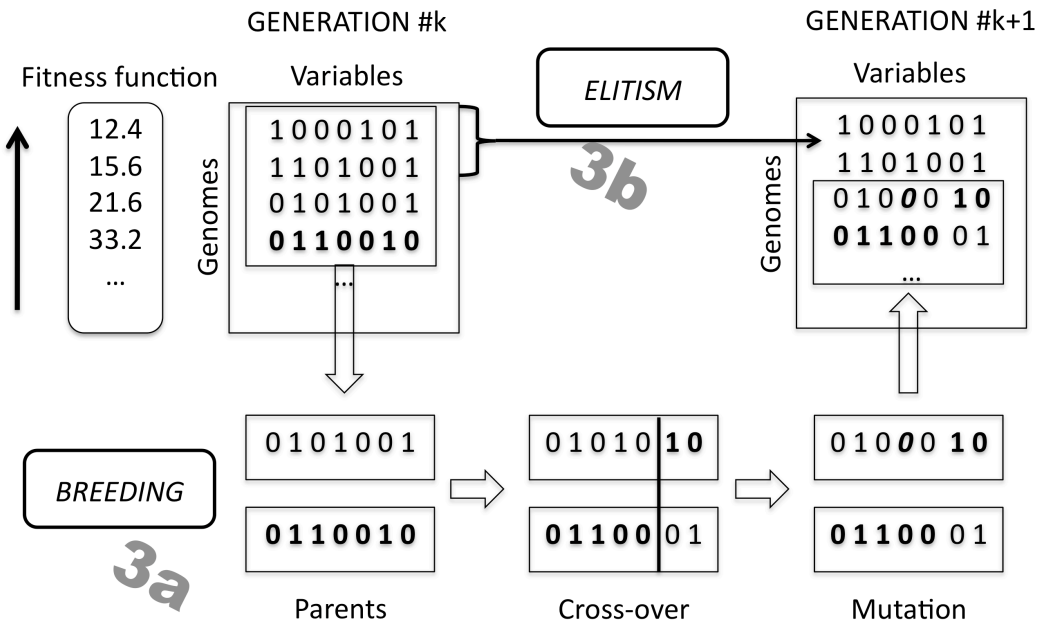


Figure 6.3: *Genotypes in the k^{th} population are sorted by descending log-likelihood on validation set. The first 10% of genotypes are directly passed down to the next generation (Elitism, 3b). The first 45% are bred: (3a-i) genotypes are randomly paired up; (3a-ii) pairs of parents are crossed-over at random positions to create 2 offspring; finally (3a-iii), a random 20% of children's genotypes are mutated (bits are flipped). Eventually, the $(k+1)^{th}$ generation is composed of 10% of the best genotypes from the previous population and 90% mutated offspring.*

The parameters in this section were initially set to values previously used in literature [308] and subsequently tuned. However, the algorithm was not found to be sensitive to them.

6.2.3.3 Stopping criterion

The maximum number of generations was set to $K = 200$ and, in order to prevent the selected variables from overfitting to the splits of the data chosen in the bootstrap procedure, an early stopping criterion was defined from consecutive log-likelihood estimates l_i and l_j as:

$$\frac{l_{j+10} - l_j}{l_{i+10}} < 0.001 \tag{6.1}$$

6.3 Different approaches to GA optimization

6.3.1 Feature selection for logistic regression

6.3.1.1 Rationale

In section 4.1.3 we have seen the added value of a feature selection technique being carried out prior to any model generation. In particular the LASSO model offers an efficient wrapping technique that optimizes the selected features as well as their coefficients. In practice however it requires the identification of a parameter α that was introduced in equation 4.25. This parameter essentially defines the amount of shrinkage to be applied to the model, which is equal to a LSE when $\alpha = 0$. The identification of this parameter usually involves a unidimensional line-search that optimizes cost function (log-likelihood) with or without the use of a cross-validation procedure. All together, this introduces additional parameters: choice of a cost function, use and type of cross-validation scheme as well as its parameters. All of these potentially reduce the potential repeatability of the study.

Alternatively, adequate implementations of the GA could iteratively identify the sub-

set of variables that maximizes the probability of the dataset given the proposed model (log-likelihood). This approach offers two important advantages: first, the GA can optimize the log-likelihood rather than a parsimonious version of the LSE as in LASSO, which may be found to be more sensitive to outliers; second, no hyperparameter (α) is required because the number of selected variables is automatically driven solely by the variation in out-of-sample performance. Consequently, the GA with such a fitness function should provide results that are at least as good as that provided by LASSO. This first implementation of the GA therefore has two objectives: to validate the GA implementation and to estimate the potential benefits of this approach compared with the LASSO.

6.3.1.2 Fitness function for logistic regression

The fitness function is evaluated during two randomly selected 8-fold validation procedures. At the end of this procedure 16 different models including exactly the same subset of features will be evaluated using different training data. Each observation is used exactly 14 times for training and exactly twice for validation. Besides, variations of these parameters (number of folds and repetitions) were not found to influence the performance even though they had a direct impact on the computational cost (this will be further discussed later). All genomes from the same generation are evaluated on the same folds, which are randomly sampled at each generation. This is meant to prevent the overfitting to some specific data split with respect to the selected variables. Finally, for each training set b in the bootstrap procedure, the performance of the subset of variables is estimated as follows:

1. logistic regression $\beta^{(b)}$ parameters are fitted with equation 3.4 to the training set $X_{training}$;
2. probability of death is then estimated on the validation set $X^{(b)}_{validation}$ with $\hat{y}^{(b)} = \pi(\beta^{(b)}, X^{(b)}_{val})$ where π denotes here the logistic function defined in equation 3.2;
3. finally, the performance of the genotype on this validation data is taken as the

log-likelihood described in equation 3.4

Finally, the score given to the j^{th} subset of variables is computed from all validation values:

$$\log\mathcal{L}(\beta|X^{(j)}) = \frac{1}{B}\sum_{b=1}^{B}\left(\sum_{i\in validation^{(b)}}(1-y_i^{(b)})(1-\pi(\beta,\mathbf{x_{ij}^{(b)}}))+y_i^{(b)}\pi(\beta,\mathbf{x_ij^{(b)}})\right) \quad (6.2)$$

6.3.1.3 Types of models

Some variables are expected to be very similar: minimum and maximum values of a feature over some period of time, chronic health conditions extracted with different techniques (ICD-9 versus NLP extraction from discharge summaries). Consequently, it is expected that the algorithm will select varying solutions across different runs. In addition to the stochastic nature of the algorithm, the selection of random folds for the evaluation of each genotype introduces another layer of variability. To capture this variability and estimate the consistency, the GA will be run 20 times. At the end of each run, the best individual from the last generation will be kept. From these runs, 4 different models will be extracted and evaluated:

- Best: the model that offers the best cross-validated performance of all genotypes;
- Averaged: the next three models are based on a variable importance metric, simply defined by the number of times a feature is included at the end of a run; as such, the GA transforms into a *filter* feature selection technique. The best ranking features are included in models of varying dimensions:
  - CV: the features are included one-by-one and a five-fold validation procedure is used to estimate the out-of-sample performance (log-likelihood) for selection of the optimal size;
  - Max: the maximum size of all selected genotypes (the best ranking in the last generation) is used;
  - Median: the median size of all selected genotypes is used.

6.3.2 Parsimonious model inspired by automatic relevance determination

Despite its interesting properties, the aforementioned GA framework for the optimization of log-likelihood during feature selection has a drawback: the different levels of cross-validation required to estimate the out-of-sample performance and prevent over-fitting. These introduce a source of variability in the results and impair the overall repeatability. Another disadvantage is the fact that each cross-validation layer further reduces the available data left for training. Ideally this should be avoided for such a data-constrained problem. In addition to this, each cross-validation layer multiplies the number of models to be fitted for the estimation of a single genotype, thereby dramatically increasing the computational cost. These are an incentives to implement a cost-function offering parsimony on a single fold.

Bayesian statistics provide an excellent framework for model comparison, which has been detailed at length by MacKay [186], and in particular in his work on ARD [188, 187]. The model comparison is achieved using a two-level inference framework: the first defines the most likely parameters given the data, while the second infers the most likely model (architecture) given the data.

Let us first introduce the posterior probability of a model's assumptions, $\mathcal{H}$, given the data D:

$$P(\mathcal{H}|D) \propto P(D|\mathcal{H})P(\mathcal{H}) \tag{6.3}$$

The *Evidence* for the data $P(D|\mathcal{H})$ in equation 6.3 can be marginalized as follows:

$$P(D|\mathcal{H}) = \int P(D|\beta, \mathcal{H})P(\beta|\mathcal{H})d\beta \tag{6.4}$$

where β are the parameters of our model. Assuming that the posterior $P(\beta|D, \mathcal{H}) \propto P(D|\beta, \mathcal{H})P(\beta|\mathcal{H})$ has a strong peak at the most probable parameters β_{MP}, we can estimate the evidence by the height of the integrand's peak times its width $\sigma_{\beta|D}$ (see figures

in MacKay et al. [188]):

$$P(D|\mathcal{H}) = P(D|\beta_{MP}, \mathcal{H}) \times P(\beta_{MP}|\mathcal{H}) \times \sigma_{\beta|D} \tag{6.5}$$

The first term of equation 6.5 is the best fit likelihood that can be estimated from equation 6.2. The second term only depends on the probability density function for the parameters (i.e. the "priors", our initial hypothesis on the model). The width of the parameters given the data (the posterior $\sigma_{\beta|D}$), can be estimated from the Hessian A (or the inverse of the covariance matrix) as $\sigma_{\beta|D} = \det^{-1/2}(A/2\pi)$ [188] where $\det(A)$ denotes the determinant of matrix A. We implemented the following fitness function on the GA, where l_j is the fitness function defined in equation 6.2:

$$fitness_j = l_j + log\left(\prod_{l=1}^{p_{kj}} P(\beta_l|\mathcal{H}) \times \frac{1}{\sqrt{\det(A/2\pi)}}\right) \tag{6.6}$$

Equation 6.6 will favour simple models with good generalization over complex models at equivalent performance, which naturally embodies the concept of Occam's razor [36]. More precisely, a model including a high number of variables will probably overfit and show a high log-likelihood l_j that will be tempered by a lower posterior of the parameters $\sigma_{\beta|D} = \frac{1}{\sqrt{\det(A/2\pi)}}$. At the same time, between two models including the same number of features the ARD will favour the model best optimizing the log-likelihood of the model, the prior of the parameters, and the posteriors $\sigma_{\beta|D}$.

The main advantage of this technique is that the whole of the available data (training and validation sets) can be dedicated to variable selection without the need for any cross-validation procedure. This means that there is a single solution to a well-defined problem. However, it is most likely that the heuristic will not consistently identify this unique solution given the large dimension of the searching space and the model will be run 20 times to capture this variability. At the end of each run, the best individual from the last generation will be kept and three different models can be built as in the previous section:

- Best: the first model is the one that offers the best cross-validated performance of all of the best genotypes;
- Mode: the exact subset of features that is most often selected (if any);
- CV, Max, and, Median are defined as in the previous model.

6.3.3 Feature selection and parameter pruning for support vector machines

6.3.3.1 Previous work on SVM parameter optimization and feature selection

SVMs were introduced in section 4.1.2.1. The optimization of the hyperparameters for SVM (kernel type and coefficient, dimension) is still an active field of research. Grid search [138, 296] is very commonly usually used for the optimization of the RBF parameters but suffers from two obvious drawbacks: if the grid interval is too small then the search is computationally too intense and if the interval is too coarse then the optimization is likely to be suboptimal. In addition to this, there is no guarantee that the identified hyperparameters remain optimal for different subsets of features, which constitutes a rationale for simultaneous feature selection and the parameter optimization. GAs have previously been used for SVM parameter optimization [216, 306] and feature selection [181, 95]. To the best of our knowledge, the first attempt to simultaneously optimize RBF parameters and select features was made by Lin et al. [183] who used Simulated Annealing (SA). GAs have also successfully been applied to this problem [140] showing better performance over the traditional grid search. However, the fitness function used in this work required the choice of new parameters to prevent overfitting. On the contrary, the approach chosen here is to rely on cross-validation strategies to adequately tune model complexity (number of features included, kernels coefficients and slack variable). Previous results presented in section 4.3.2 showed that the combination of C-SVM with RBF kernel seemed optimal, both in terms of computational cost and performance. Consequently, this is the architecture that was chosen to be optimized here.

6.3.3.2 Description of the fitness function

The slack variable C and the RBF σ parameters were coded on nine bits with the τ function defined as follows

$$\tau : \pi \rightarrow 2^{\frac{\text{bin2dec}(\pi_m)}{50} - 10} \tag{6.7}$$

so that all possible values were log-distributed between 2^{-9} and 2^{0} which conveniently covers the area of best performance for both parameters according to the grid-search carried out in section 4.1.2.1 while keeping away from computationally expensive areas of the search-space also identified in that section.

The initialisation of the population was similar to that previously described for the part of the genotype that relates to feature selection. The 18-bit-long coding for the hyperparameters was simply initialized with a random Bernoulli distribution with the mean arbitrarily set to $\mu = 0.3$; this parameter was not found to be sensitive. After initialization, the evolution process was carried out as detailed above, independently of the segment type (variable or parameter). Finally, the genotype evaluation (combination of subset of variables and hyperparameters) was carried out with the fitness function presented in equation 6.2. Unlike previous studies, the GA was only run once; the best performing individual in the last generation was selected and the model was finally fitted to the entire training set using features and parameters indicated by this genotype.

6.4 Results

The performance achieved by the different implementations of the GA is presented in table 6.2, which reports similar metrics of performance as seen in the previous chapters and on the very same folds. A column has been added to the table in order to indicate the size (number of included variables) of each model. This is meant to show the variations in model size with respect to model performance, which are implicitly included in the ARD fitness function.

The best model performance as defined by the Nll was reached by the Automatic

Table 6.2: *Results provided by different implementation of the GA sorted by decreasing performance: Automatic relevance determination (ARD), Support vector machine (SVM), and Logistic regression (LR) .*

Fitness function	Model selection	Model size	Nll	AUROC	SMR
ARD	MaxSize	34	333.1±8.2	80.0±0.6	0.99±0.10
ARD	MedSize	26	334.8±7.0	79.4±0.7	0.98±0.11
SVM	Best	100	337.9±6.1	80.0±0.3	0.99±0.09
ARD	Cross-validated	60	337.5±15.4	80.3±1.0	0.96±0.12
ARD	Best	32	341.0±7.6	79.1±0.6	0.96±0.11
ARD	mode	13	341.1±7.0	78.0±0.9	0.95±0.11
LR	Cross-validated	69	341.8±10.8	80.7±1.1	0.97±0.11
LR	MaxSize	40	352.9±13.1	80.6±0.1	0.99±0.07
LR	MedSize	36	354.5±12.4	80.3±0.1	0.98±0.08
LR	BestModel	40	357.2±32.4	79.1±2.2	0.97±0.05

Relevance Determination (ARD) model of larger size amongst all 50 runs of the GA. This model had a Nll of 333.1 ± 8.2 and an AUROC of 80.0 ± 0.7% with a model size of $k = 34$. This model did not offer a statistically significant difference to the other top five models of this table with any of the metrics presented in section 3.1.2. Interestingly, the model with the second best performance only included $k = 26$ variables and showed good performance $Nll = 334.8 \pm 7.0$. The model based on the logistic regression fitness function seemed less likely to have generated the data according to the likelihood values. However, the model of the largest size ($k = 66$), whose dimension was identified thanks to an additional level of cross-validation, offered the best discriminatory power ($AUC = 80.8 \pm 1.1$). The SVM-based model did not offer the best performance ($Nll = 337.9 \pm 6.1$, $AUC = 80.0 \pm 0.3$) but compared advantageously to previous occurrences of a similar SVM model (C-SVM with RBF) trained on the same subset of variables, as presented in section 4.3.2 ($Nll = 377.0 \pm 3.4$, $AUC = 74.3 \pm 0.8$).

Figure 6.4 illustrates the convergence of the different versions of the GA across all runs. While both LR and SVM (in black and green, respectively) are estimated on 2×8 folds in order to prevent overfitting, the ARD (in blue) is estimated on the same unique

training set, which is why the blue dots show a clear positive offset in performance. The difference in performance between the figures seen in this figure and those in table 6.2 is due to the process of external validation: the cross-validated results in green and black in figure 6.4 are indeed closer to the real performance on unseen data than the results generated on the training set (blue dots).

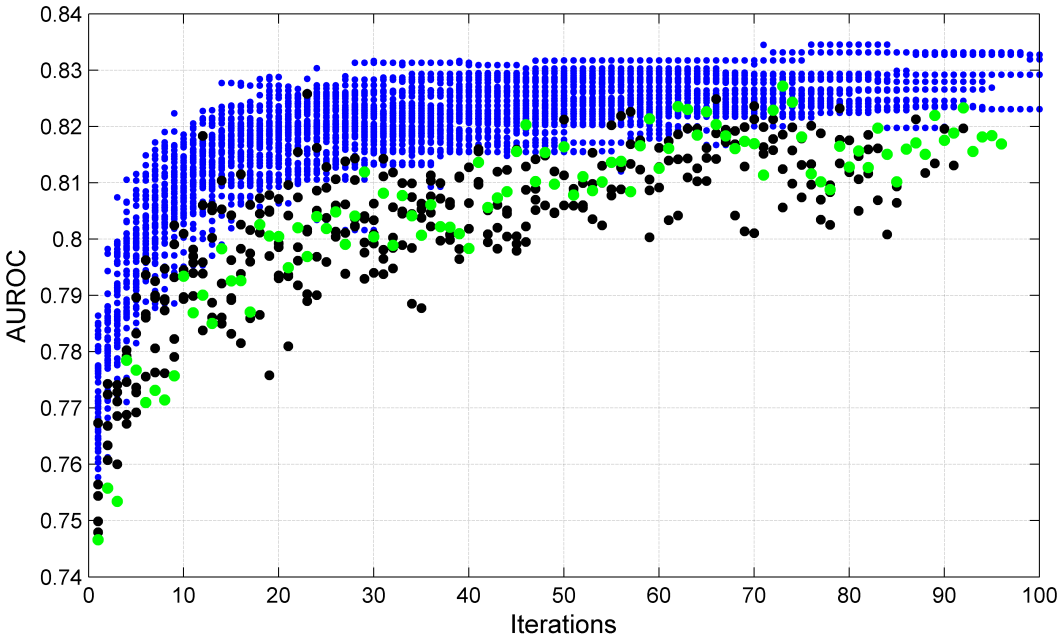


Figure 6.4: *This figure shows the evolution of the fitness function over the generations (x-axis) for the logistic regression (LR, black), automatic relevance determination (ARD, blue), and the support vector machine (SVM, green).*

The computational cost of the different approaches is presented in table 6.3 that shows for each technique, the number of runs planned, and the average number of iterations per run. The worst technique in terms of computational cost was the SVM-based fitness function that required more than 20 minutes per iteration amounting to more than 3 hours per run and per fold. Second, the LR-based fitness function was estimated to take 20.8 ± 7.2s per iteration, which amounted to an estimated 60 minutes per run on each fold. Finally, the ARD-based fitness function was found to be the fastest technique with an average iteration taking roughly 1 second making it faster than the SVM approach by three orders of magnitude.

Finally, figure 6.5 shows the iterative feature selection process averaged across all

Table 6.3: *This table presents the computational time for different versions of the GA with the number of independent runs, the average number of iterations per run and the associated time in seconds. All results are presented with the median and quartiles when relevant.*

GA type	Runs (#)	Iterations (#)	Time per iteration (s)
LR	5	82 (44 – 87)	20.8 (18.7 – 26.9)
ARD	50	67 (54 – 84)	1.0 (0.8 – 1.2)
SVM	1	96 (N.A.)	1207.4 (N.A.)

runs for the LR and ARD genetic algorithm.

6.5 Discussion

Intuitively, one can assume that the performance of a GA largely depends on the chosen fitness function. Different techniques optimizing the exact same criteria, should reasonably provide an equivalent solution. For instance, it is expected that a GA using a fitness function based on Mutual Information would not outperform Joint Mutual Information (JMI) based techniques [307]. In this chapter it was hypothesized that the Nll cost function optimized by the GA with a logistic regression model may offer several advantages over LASSO. First, the Nll may prove more robust to outliers than a penalized version of the least-square cost function implemented in LASSO. Then, the automatic determination of the model dimensionality from the out-of-sample performance removes the need for an additional hyperparameter to quantify the amount of shrinkage, which may have some additional benefits. Despite a slight gain in performance provided by the GA approach for all performance metrics ($Nll_{LASSO} = 337.8 \pm 6.7$ and $Nll_{GA} = 341.8 \pm 10.8$), no statistically significant difference could be established ($IDI = 0.03$ and $p_{IDI} = 0.32$). Yet, this result confirms the successful optimization of the fitness function with the GA (results are equivalent to those of LASSO) and consequently gives a basis for further improvements to the GA framework. The positive results given by the ARD version of the GA are also encouraging even though the modest gain in performance may not neces-

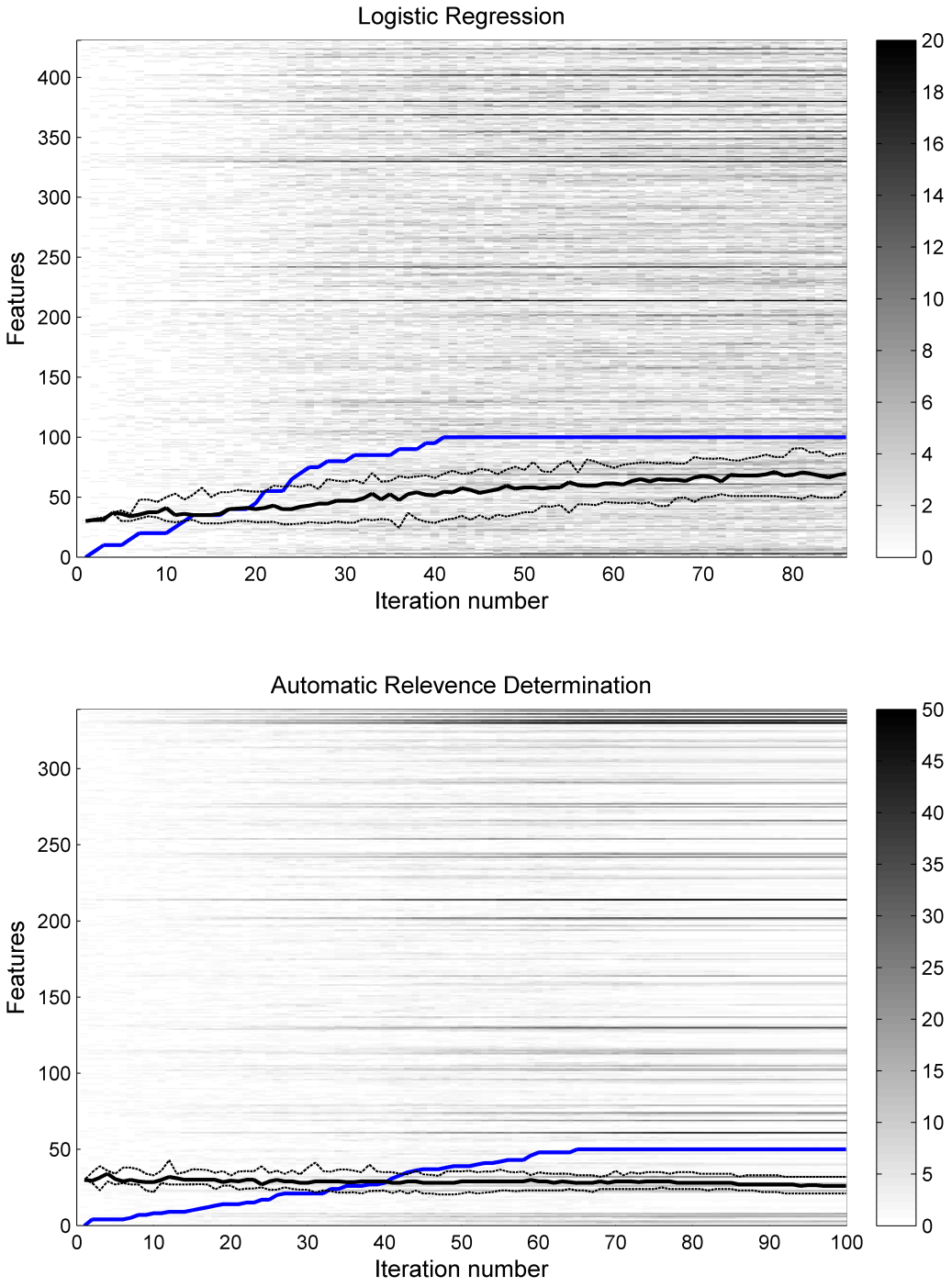


Figure 6.5: *This figures plots the average evolutions of genomes during consecutive iterations of the algorithm. In the background is shown the average number of selections in shades of grey with the corresponding colour map on the right-hand side of the figure. The total number of iterations depended on the activation of an early-stop criterion and all GA runs were therefore aligned to the last generation on the right of the plot. The blue line represents the number of GA runs considered at every iteration (x-axis) for the averaging starting from 1 on the left to the total number of runs on the right. The black line represents the average number of selected variables (y-axis) at the given generation with the 25^{th} and 75^{th} percentiles shown as dashed lines.*

sarily justify the added computational cost when compared to the LASSO.

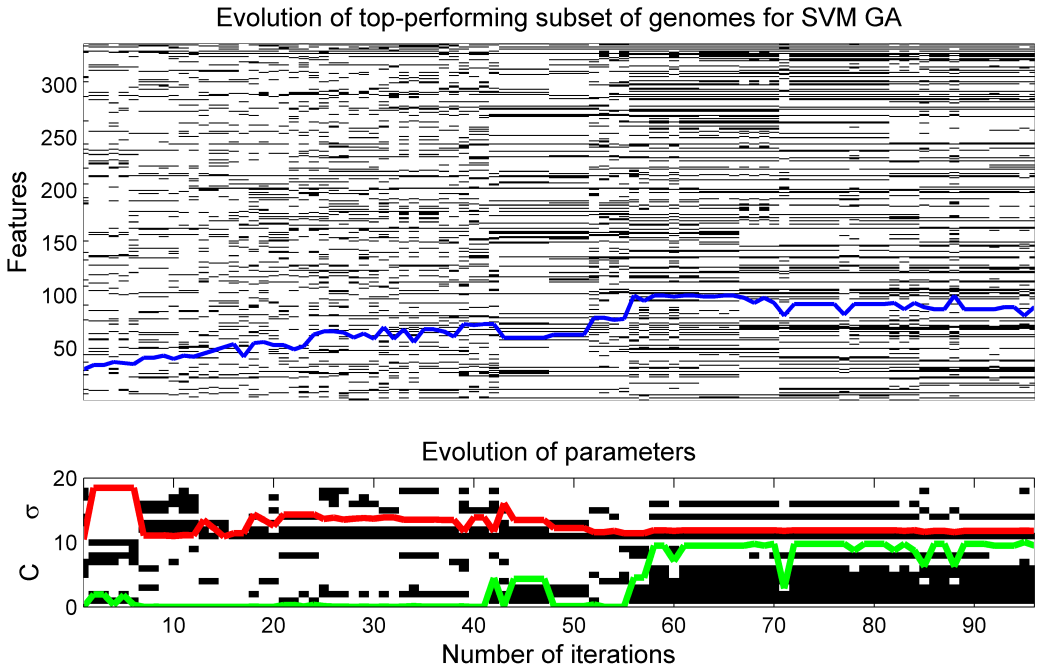


Figure 6.6: *Evolution of genomes across generations. (TOP) The best genome of each generation is binary coded (black for selected and white for not-selected) and the resulting model size is indicated in blue. The y-axis therefore indicates the feature number (for binary coded selection) and the number of features included in the model. (BOTTOM) Evolution of genes coding for the hyperparameters from 1 to 8 for C and 9 to 19 for σ together with their respective interpreted values in green and red, respectively.*

Traditional SVM modelling techniques involve a *filter* feature selection technique followed by a rudimentary grid-search for the identification of hyperparameters. In this chapter it was hypothesized that the joint optimization of a feature subset and hyperparameters (C and γ) would increase model performance. This was confirmed to a limited extent on this dataset since all metrics consistently improved with this framework even though no statistical significance could be observed. To improve the understanding of this phenomenon, the evolution of the best feature subset and parameters across different generations is plotted in figure 6.6. In this figure, the y-axis indicates the feature number (for binary coded selection) and the number of features included in the model (for the model size in blue). The convergence for the feature selection process is indicated by the stratified aspect of the second-half of the plot (right) from the 50^{th} iteration onwards while the blue line (model dimension) levels out. The optimal model dimension-

ality selected by the GA without the need for an additional hyperparameter was therefore slightly below 100 features. The bottom of the figure shows the simultaneous evolution of the genes binary coding for the hyperparameters and their interpreted values (σ in red and C in green). The joint observation of both halves of this figure reveals abrupt changes in the values of parameters at each important modification of the feature subset. This phenomenon supports the joint optimization of the feature subset and the SVM hyperparameters.

The comparison of different computational costs presented in table 6.3 clearly discourages the use of the SVM-based fitness function. The dramatic increase in processing time is not balanced by any improvement in model fit or discriminatory power. On the contrary, the ARD-based fitness function offers the best model fit of all and jointly provides the lowest computational cost. Clearly, the use of any cross-validation procedure implemented to constrain model size and prevent overfitting certainly impairs speed. For instance, LR and SVM-based fitness functions require every model to be fitted and applied 5×10 times in order to estimate out-of-sample performance. A second factor of importance could be the convergence properties of a given fitness function that are influenced by the activation of the early-stopping criterion presented in section 6.2.3.3. From the data presented in table 6.3 however, no statistically significant variation of iteration number was found between the techniques with the use of a non-parametric Wilcoxon rank-sum test (for all pair-wise comparisons $p > 0.1$). Finally, the difference in computational cost may also be explained by various additional parameters: training set size, the number of features, and the type and performance of optimization technique (iterative versus closed form).

Interestingly, the simplest model was a logistic regression model including only 13 variables in a prediction rule described by equation 6.8 below (all coefficients were found to be statistically significant with $p < 0.01$). These are the most significant clinical covariates according to the optimization rule (ARD, equation 6.6) and therefore deserve special attention. First of all, existing severity scores (SOFA, APS, and Van Walraven) account for

a large part of the selected variables. On the one hand, this reveals the positive contribution of score aggregation, the coefficients of which were derived from external databases. On the other hand, this impairs the apparent simplicity of the model by requiring the collection of additional clinical covariates. Then, the absence of mechanical ventilation and the amount of lactate Ringers administered to the patient were both found to be negatively related to the adverse outcome, suggesting (i) a beneficial effect of this treatment or (ii) the presence of strong confounding factors potentially related to the baseline risk of the populations to whom such treatments are given. For instance, lactate ringers could be preferably administered to cardiac ICU patients for whom the overall mortality rate is much lower (see table 2.4). Similarly, two CHCs (pulmonary failure and depression) were found to be negatively related to the outcome, possibly for equivalent reasons. Surprisingly, $Ethnicity$ was selected in its categorical version rather than the dummy-coded one ($EthnicityIsBlack$, $EthnicityIsCaucasian$, ...), while the categories were attributed randomly as follows: African (1), Caucasian (2), Asian (3), Hispanic (4), other (5). The weight identified for this categorical variable was consistently positive suggesting that different ethnicities react to sepsis-induced hypotension in such a way that severity is somehow related to the random ordering given above. However, because social status is not coded in the database and is closely related to some ethnicities in the US, the interpretation of such a finding is difficult. Having said this, it was earlier reported that patients of African descent may have an overall lower severity of sepsis [6], which seems to corroborate this result. In terms of physiological variables, unsurprisingly the minimum lactate value, the total urine volume, the minimum respiratory rate, and the mean temperature over the first 24 hours following admission to the ICU were also selected as strong predictors of mortality. Less obvious is the presence of the WBC standard deviation (σ) in this score: the greater the variation the healthier the patient.

$$\begin{aligned}\pi(x) = &-1.20 + 0.46 \times \text{SOFA}_{\text{First}} + 0.37 \times \text{Lactate}_{\text{Min}} - 0.37 \times \text{Urine}_{\sum} \\ &+ 0.35 \times \text{APS}_{\text{First}} + 0.32 \times \text{Mech. Vent.}_{\sum} + 0.29 \times \text{Van Walraven}_{\text{First}} \\ &+ 0.29 \times \text{Resp. Rate}_{\text{Min}} - 0.27 \times \text{Ringers}_{\sum} - 0.26 \times \text{Temp.}_{\mu} \\ &- 0.25 \times \text{Depression} - 0.20 \times \text{Surgical Unit} - 0.20 \times \text{Pulmonary} \\ &+ 0.20 \times \text{Ethnicity} - 0.19 \times \text{WBC}_{\sigma} \end{aligned} \tag{6.8}$$

Finally, additional benefits of the GA approach may also include the optimization of particular fitness functions for which:

1. no analytical solution can be found or at the cost of constraining assumptions such as one on the type of distribution for the parameters or variables (Gaussian or normal hypothesis);

2. where gradient descent or sequential elimination techniques are likely to fail given: 1) the large size of the search space and 2) the presence of local minima.

Conclusion

In summary we found marginal but convincing evidence of the potential of GA frameworks for feature selection subsequent to parameter pruning. The SVM-based GA model did not seem to provide superior performance but compared well to earlier instances of SVM models (filter feature selection and grid search) suggesting a potential benefit that may be demonstrated on a larger database. Having said this, the large computational cost associated with this modest gain in accuracy did not seem favourable. Finally, the most interesting implementation of the GA was that using an ARD-inspired fitness function. In fact, there is a clear rationale for optimizing a criteria providing a parsimonious model without the need for additional cross-validation layers. This decreases the computational cost and discards the associated source of variability in the results.

Chapter 7

Adding dynamic information from clinical data to the prediction of mortality

Introduction

We have seen in previous chapters that different levels of customization improve the accuracy of outcome prediction: recalibration; adjustment of covariate weights; inclusion of more representative covariates; and potentially the use of more advanced modelling techniques, including heuristic search algorithms such as genetic algorithms. These improvements have raised the performance of the predictive rule to improved calibration levels ($p_{HL} > 0.05$) and acceptable discriminatory levels ($AUROC > 80\%$). Yet it is clear that these rules are still not good enough to be applied at the individual level [293]. In the absence of larger cohorts or additional covariates, alternative approaches to data modelling need to be investigated. While the purpose of this work is to estimate patients' severity, we hypothesize that the dynamic information contained in the physiological trends may bring additional information.

Dynamic comes from "power" in Greek ($\delta\upsilon\nu\alpha\mu\iota\sigma$, dunamis) and is used today by

physical scientists to describe "a process or a system characterized by constant change, activity, or progress". According to this definition, nature is essentially *dynamic* and physiology (the study of vital functions) is one of its most complicated mechanisms characterised by constant change, activity, or progress witnessed at every level of life, from the variation of ionic concentrations in a cell to the size of species populations.

Certainly the study of dynamics should reflect vital functions and may correspond to the severity of a disease. For instance, the evolution of a patient is condition over time may contain valuable information: two patients with a mortality score of x on day 2 after admission might have followed different paths (recovery versus disease progression), which may relate to actual severity. In addition to this, we hypothesize that these trends may be of particular interest with respect to *endogenous* and *exogenous* impulses. An endogenous impulse is defined here as a sudden and important change in physiology: cardiac arrest and shock (as defined in section 1.4) are two good examples of this. The drop in blood pressure occurring during shock is believed to trigger a series of physiological reactions, including the acceleration of heart rate and respiration via activation of the sympathetic system. Capturing this reaction may relate to the patient's overall physiology and therefore improve any estimation of severity. Similarly, an exogenous impulse is defined by a sudden and important physiological change induced by an external intervention. Every intervention in the ICU potentially meets this definition and different degrees of response can usually be observed, we believe, in relation to some outcome. For instance, patients who are challenged with fluids have a typical haemodynamic response (increased blood pressure) that potentially vanishes while the patient deteriorates. Capturing the degree of a patient's response to the intervention will certainly reflect some aspects of their physiology that are not necessarily accounted for in traditional models of severity. To conclude, the change in approach suggested in this chapter promotes dynamic (rather than static) modelling to capture variations in severity over time, possibly with respect to specific internal and external events. Interestingly, septic shock offers an ideal framework for the exploration of such a hypothesis. Indeed,

both the internal events (hypotension and shock) and the treatments (vasopressors and fluid) occur in a well-defined sequence that triggers highly standardized protocols for the management of the condition. This hypothesis is therefore tested on the data set presented in this work and the results are compared with the approaches presented previously.

7.1 Exploring the added value of dynamic information to improve the prediction of mortality

7.1.1 Comparison technique

In order to explore the potential benefit of dynamic information for the prediction of mortality, different datasets can be extracted and processed. The same machine learning techniques (for extraction of feature importance and prediction rule) can be applied to these different datasets, including those presumably capturing variations in patient physiology over time. Finally, the resulting performance metric associated with each dataset was assumed to reflect the predictive power captured by it as it is described below.

The choice of machine learning technique used for comparison of the datasets is based on both theoretical and practical considerations, leading to the use of LASSO. First of all, the LASSO benefits from a strong theoretical background, which provides strong arguments in favour of the validity of the technique. Last but not least, the much greater computational times of SVM-based models, BEF, and GAs also support the use of the LASSO for this comparison.

Consequently, the baseline for all comparisons is the LASSO model presented in section 5.5.2 that was derived from the extended set of features collected during the first 24 hours following admission to the ICU. The comparison with the new models can be carried out with the statistical techniques presented in section 3.1.2, notably the IDI. Very

specific care is given to the time line of the analysis ensuring that the studies compare information derived from exactly the same period of time with respect to the hypotensive episode for every patient. The relation between the time of the prediction and the model performance will then be developed in section 7.2.1 below.

7.1.2 Description of the datasets

The hypothesis presented in this chapter is two fold: (1) the evolution of patients' physiology may estimate their severity in a more accurate way than static data; (2) the patient's response to endogenous and exogenous physiological impulses may improve this prediction further. The first hypothesis can easily be explored within the traditional time window (the first 24 hours), even though the typical sampling rates of some clinical variables (once a day) may require a slightly larger time window (first two days). The second hypothesis however needs the analysis to be carried out at the time of physiological derangement, which requires the static comparison of data extracted over the same period of time with respect to the hypotensive episode.

Figure 7.1 represents the different analyses presented in this chapter. The first study evaluates the benefit of splitting the traditional first 24-hour time-window (Model $D1$) into two 12-hour segments (Model $D1 - \Delta_{12}$) in order to capture patients' physiological trends over this time window and thereby estimate the potential additional predictive value. Then, to account for relatively low sampling rates (typically once a day), the study is repeated using two segments of 24 hours each (Model $D1D2 - \Delta_{24}$) which are compared to the first 2 days as a single block of data (Model $D1D2$). The second study investigates more specifically data surrounding the hypotensive episodes. To do this, three time-windows are considered: the data before the hypotensive episode, during it, and afterwards, from which different models are computed as detailed on figure 7.1: $HE - PRE$, $HE - POST$, $HE - OVER$, and $HE - \Delta_\tau$. Different window length surrounding the hypotensive episode are finally explored: $\tau = 12$, 16, 20, and 24 hours.

On each time-window and for each covariate, data statistics are extracted such as

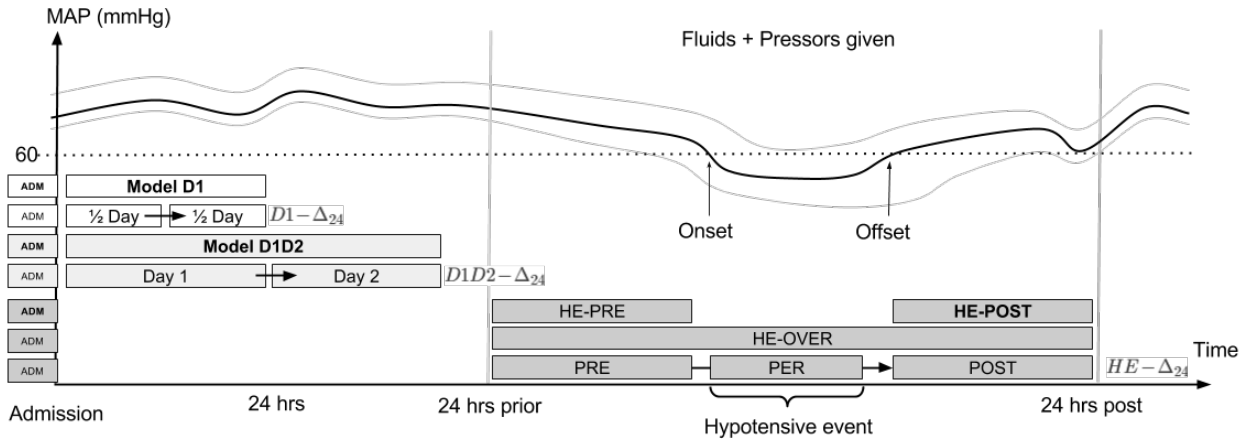


Figure 7.1: *An example of arterial blood pressure (in plain: ABP_{Mean} and dashed: $ABP_{Diastolic}$ and $ABP_{Systolic}$ in dashed) time-series from admission showing an occurrence of a hypotensive episode (HE). With respect to this event as well as admission time, different epochs are extracted to estimate the benefit of dynamic information. In shades of grey (white, pale, dark) are seen three different studies with the baseline models printed in bold: $D1$, $D1D2$, and $HE-OVER$. The inclusion of dynamic information is represented by a black arrow.*

minimum, median, maximum value, and standard deviation in a similar way to that presented in the previous chapters. For the first two studies (Models $DX - \Delta_\tau$), dynamic information is simply extracted by looking at the difference between the median of the first and second time window. For instance, the evolution of the heart rate over the hypotensive episode is considered to be:

$$\Delta HR = \overline{HR}(12 \leq t < 24) - \overline{HR}(0 \leq t < 12) \tag{7.1}$$

where $\overline{HR}(t_0 \leq t < t_1)$ denotes the median value of x between t_0 and t_1. For the second study, dynamic information will be incorporated in a slightly more complex way to account for the variation in length of the hypotensive episode. More precisely for a given variable (HR), the slope defined by the closest available sample on each side of the

hypotensive episode is used to characterize the dynamic response as follows:

$$\begin{aligned}\text{Post} =& \text{First}\left[HR(t_{offset} \geq t > t_{offset} + \tau)\right] \\ \text{Pre} =& \text{Last}\left[HR(t_{onset} - \tau > t \geq t_{onset})\right]\end{aligned}$$

$$\frac{\Delta HR}{\Delta t} = \frac{\text{Post} - \text{Pre}}{t_{offset \rightarrow \text{First}} - t_{\text{Last} \rightarrow onset}} \tag{7.2}$$

where "Post" denotes the first chronologically available value after t_0 or Not A Number (NaN) if not found before $t_0 + \tau$ with $\tau = 24$ hours; $t_{0 \rightarrow \text{First}}$ is the associated time-stamp. Similarly, "Pre" denotes the closest available value before t_0 (NaN if not found after $t_0 - \tau$) and $t_{\text{Last} \rightarrow 0}$ the associated time-stamp. With the definition in equation 7.2, dynamic information is simply incorporated with the gradient of the variable (here HR) calculated from the start to the end of the hypotensive episode.

Finally, the sizes of the different feature subsets included for the analysis are presented in table 7.1 showing dimensions that are substantially larger than those seen in previous chapters. This table only displays unique features; for instance minimum, median, and maximum values computed over 24 hours for a feature typically sampled daily were discarded. Similarly, features from which more than 99.95% of observations had identical values were automatically removed from the analysis (this corresponds to 10 unique observations on the $2,143$ patients in this dataset). These two rules were enforced because: (1) redundant features are likely to harm the feature selection process and (2) the different levels of cross-validation impose a minimum number of values available per feature for the correct identification of variable weights. After such filters, the feature subset for the dynamic model $HE - \Delta_\tau$ still included more than one thousand distinct features.

Table 7.1: *List of the different models evaluated showing model size (number of free parameters) as well as the relative contribution of different temporal segments: admission, before (PRE), during (PER), and after (POST) the hypotensive episode. DELTA denotes the variable meant to capture the dynamic information as detailed in section 7.1.2. The numbers given in this table indicate how many variables are considered prior to model design (before feature selection); other features showing perfect redundancy are discarded. The numbers in brackets indicate how many of these are coding for the absence of a value (missing value binary flag).*

Model	Admission	PRE	PER	POST	DELTA	TOTAL
HE-PRE	140 (40)	330 (162)				470 (202)
HE-POST	140 (40)			308 (146)		448 (186)
HE-OVER	140 (40)		388 (177)			528 (217)
HE-Δ_τ	140 (40)	330 (162)	72 (28)	308 (146)	159 (80)	1009 (456)

7.1.3 Results

The comparison of the performance of the different models is presented in table 7.2. First and foremost, the dynamic models over the first day ($D1 - \Delta_{12}$) and first two days ($D1D2 - \Delta_{24}$) clearly outperform static models over the same time period ($D1$ and $D1D2$, respectively), showing a statistically significant difference ($IDI = 3.51$ and $IDI = 2.46$ with both $p_{IDI} < 0.05$). The second study however could not reveal any statistically significant difference to support the superiority of the dynamic approach (Model $HE - \Delta_{24}$) when compared to modelling on data extracted from the same time window (Model $HE - OVER$). Yet these two models incorporating data surrounding the hypotensive episode ($HE - OVER$ and $HE - \Delta$) did significantly better than a model considering only data after the offset of the hypotensive episode ($IDI = 1.5$ and $p_{IDI} = 0.04$).

The thirty most important features identified with LASSO on different subsets of variables are presented in table 7.3 for the best performing models in each study. The features selected for the "static" models are fairly consistent with what would be expected for such a population (patients with infection, SIRS, organ failure and shock) as well as with the findings reported in section 5.5.2. In the dynamic model $D1D2 - \Delta_{24}$, Temperature was included over $D1$ and $D2$ and the increase in bicarbonates (HCO_3 $[\Delta_{24}]$)

Table 7.2: *Performance of models built from different feature subsets. The final model dimension determined by the LASSO procedure on the entire dataset is shown together with different metric of cross-validated performance (3-fold): negative likelihood (Nll), area under the receiver operating curve (AUROC), and standardized mortality ratio (SMR).*

Model	Dimension	Nll	AUROC	SMR
$HE-\Delta_{24}$	145	299.2±6.2	85.6±0.9	1.03±0.08
$HE-\Delta_{20}$	101	303.9±15.7	85.3±1.2	1.02±0.10
$HE-OVER_{24}$	112	304.5±14.5	85.3±1.2	1.01±0.02
$HE-\Delta_{16}$	175	304.6±16.0	85.1±1.3	1.00±0.10
$HE-\Delta_{12}$	165	306.3±13.9	85.0±1.0	1.01±0.14
$HE-POST$	99	311.0±9.8	84.2±1.3	1.01±0.11
$D1D2-\Delta_{24}$	97	321.8±10.4	82.7±0.9	0.96±0.05
$D1-\Delta_{12}$	135	323.1±11.5	82.5±0.7	0.94±0.09
$HE-PRE$	99	326.9±31.2	83.1±1.6	0.95±0.09
$D1D2$	86	337.4±7.0	81.0±1.1	0.95±0.11
$D1$	101	347.6±13.9	80.1±1.5	0.94±0.05

directly related to an increase in severity. Similarly, in model $HE-\Delta_{24}$ the initiation of renal replacement therapy (RTT Δ_{HE}) was found to relate to a worse outcome while the identification of blood infection over the hypotensive episode (Blood infection Δ_{HE}) increased chances of survival.

The selection of variables such as "presence of bypass surgery" in the different models reveals the important contribution of cardiac surgery patients in this population. Elective surgery patients tend to have a much lower severity than other ICU patients and thereby constitute a very specific ICU population. In particular, CABG patients constitute a homogeneous group of patients admitted to the CSRU where the overall in-hospital mortality is just 8.4% (See table 2.2). We included these patients on the basis that they may develop infection and sepsis during their ICU stay as a consequence of the invasive surgical procedure. These patients constitute a tenth of the final cohort, 3.4% of all CSRU patients, and had a mortality rate of 12.3% against 31.4% in non-surgical patients. Arguably, the difference in severity between these two groups could not entirely be explained by different levels of patient monitoring and care. Therefore it has to be considered that some

Table 7.3: *Thirty most important variables selected by the LASSO for 4 models representing the dynamic model and its baseline over the same period of time.*

Model $D1D2$	Model $D1D2 - \Delta_{24}$	Model $HE - OVER$	Model $HE - \Delta_{24}$
Temperature – μ (-0.30)	Age (0.28)	Age (0.34)	Age (0.33)
Urine – $\sum$ (-0.26)	No. Met. Cancer (-0.27)	No Met. Cancer (-0.32)	No Met. Cancer (-0.31)
Age (0.23)	No Hep. Failure (-0.23)	SOFA [OVER] – μ (0.31)	SOFA [PRE] – μ (0.24)
APS First (0.23)	SAPS-I First (0.21)	No Bypass [OVER] (0.28)	Shock Index [POST] – μ (0.23)
Eye Open – μ (-0.22)	SOFA [D1] – Min (0.21)	Van Walraven (0.27)	Urine [POST] – $\sum$ (-0.23)
Lactate – Min (0.18)	No Bypass [D1] (0.15)	Urine [OVER] – $\sum$ (-0.27)	No Hep. Failure (-0.22)
No Bypass (0.17)	Elective admission (-0.15)	Eye Open [OVER] – μ (-0.26)	Eye Open [POST] – μ (-0.20)
WBC – Min (0.17)	No Cirrhosis (-0.12)	No Hep. Failure (-0.22)	Temperature [POST] – μ (-0.16)
Resp. Rate – μ (0.15)	No Mech. Vent. (-0.12)	Elective admission (-0.21)	No Ringers [PRE] (0.16)
Shock Index – μ (0.15)	APS [D2] – Max (0.12)	APS [OVER] – Max (0.20)	No Cirrhosis (-0.16)
pH [MV] (-0.14)	Age [MV] (0.12)	No Cirrhosis (-0.20)	Elective admission (-0.15)
FiO_2 – σ (-0.14)	Ethnicity (0.11)	BUN/Creat. [OVER] – Min (0.20)	SOFA [POST] – Max (0.14)
BUN – Min (0.13)	Temperature [D1] – μ (-0.11)	Temperature [OVER] – μ (-0.16)	Age [MV] (0.13)
No Intubation (-0.13)	HCO3 [Δ_{24}] (-0.11)	Age [MV] (0.16)	RTT [Δ_{HE}] (0.12)
Urine – μ (-0.11)	Surgical ICU (-0.11)	ALP [OVER] – Min (0.12)	No Graft (0.12)
No Ringers (0.11)	Urine [D1] – $\sum$ (-0.10)	Lactate [OVER] – Min (0.11)	Blood infection [Δ_{HE}] (-0.11)
Neuro SOFA – Min (0.09)	Depression (-0.10)	SpO_2 [OVER] [MV] (0.11)	SOFA [POST] – μ (0.09)
ABP_{dia} – Max (-0.09)	Glucose [D2] – μ (0.10)	No Pseudo. Monas. Aer. [OVER] (-0.11)	APS [PRE] – μ (0.08)
INR – μ (0.09)	SOFA [D1] – μ (0.09)	Depression (-0.11)	FiO_2 [POST] – Min (0.07)
Glucose – μ (0.09)	No Enterococcus Sp. [D2] (-0.09)	APS [OVER] – μ (0.10)	Ethnicity (0.07)
Blood infection – μ (0.08)	Eye Open [D2] (-0.09)	No Sedatives [OVER] (0.10)	Glucose [PRE] – σ (0.07)
Double Product – μ (0.08)	Lactate [D1] – Min (0.09)	WBC [OVER] – Min (0.10)	GCS [POST] – Max (-0.07)
SpO_2 – μ (-0.08)	pH [D1] – σ (0.08)	HE Length (0.10)	INR [POST] – μ (0.07)
INR – Max (0.08)	Shock Index [D2] – μ (0.08)	Ethnicity (0.10)	Depression (-0.07)
Hem. SOFA – Min (0.07)	HE Length (0.08)	Surgical ICU (-0.09)	MAP [POST] – Min (-0.07)
Chloride – μ (-0.07)	Temperature [D2] – μ (-0.08)	No Lymphoma (-0.09)	Surgical ICU (-0.06)
Na – Min (0.07)	Ringers [D1] – $\sum$ (-0.08)	INR [OVER] – μ (0.09)	SpO_2 [POST] – μ (-0.06)
CO – Max (-0.07)	Ringers [Δ_{24}] (0.08)	No Ringers [OVER] (0.09)	SAPS-I First (0.06)
Sedatives – $\sum$ (-0.07)	Resp. Rate [D2] – μ (0.07)	Motor Resp. [OVER] – Max (-0.09)	APS [POST] – Max (0.06)
No 5% Dextrose (0.06)	Van Walraven [D1] – μ (0.07)	FiO_2 [OVER] – Min (0.08)	ALP [POST] – Min (0.06)

of these patients might be artefacts of the method chosen for the cohort identification. First, the model $HE-\Delta_{24}$ was re-trained on all non-CSRU patients and subsequently applied to all cardiac surgery patients. Interestingly, the model showed good discriminatory power with an AUC of 82.1% suggesting that a severity model derived from non-surgical sepsis patient could adequately discriminate survivors from non-survivors in a population of surgical patients thought to have sepsis-induced hypotension. Conversely, the model could not achieve good calibration and overestimated mortality with an SMR of 1.27. A first explanation for this results is that the baseline risk is different for these two sub-populations (different mortality rates). The model using the entire population will account for this by including variables such as "presence of bypass surgery" and offer good calibration, while the model trained without these patients could not. Finally, the drop in discriminatory power observed when the model is applied to cardiac surgery patients (from 85.6% to 82.1%) may also indicate the presence of patients wrongly selected by the cohort identification technique. In order to estimate the potential impact of these patients on the model, all models described in table 7.3 were retrained using only the non-CABG patients. Table 7.4 shows the selected variables for this population. It indicates that over 90% of the top-thirty predictors present in table 7.3 are also found in this table and removed features obviously include the ones that are specific to CABG patients ("No Bypass"). To conclude, these results highlight the drawbacks of any cohort identification techniques in general and the one chosen in particular. Simultaneously, they also strengthen the degree of confidence one might have in the findings provided in this chapter by showing that a variation in the study population yield to a marginal modification of the result.

Finally, a closer look at the models' performance with respect to prediction time reveals that models developed from "late" data with respect to admission time seem to perform better. This is illustrated with the performance of the following models that were of steadily increasing accuracy: $D1$, $D1D2$, $HE-PRE$, and $HE-POST$. Yet, the superiority of $HE-\Delta_{12}$ which makes a prediction 12 hours ahead of $HE-POST$ as well

Table 7.4: *Thirty most important variables selected by the LASSO for 4 models representing the dynamic model and its baseline over the same period of time using only patients who did not undergo cardiac surgery.*

Model $D1D2$	Model $HE - \Delta_{24}$	Model $D1D2 - \Delta_{24}$	Model $HE - -OVER$
Age (-0.30)	Met. Cancer (0.30)	Age (0.31)	Age (0.35)
APS First (0.24)	Age (0.30)	Met. Cancer (0.31)	Met. Cancer (0.35)
Urine – $\sum$ (-0.24)	APS [POST] – Max (0.25)	SOFA [D1] – Min (0.29)	Van Walraven (0.32)
Eye Open – μ (-0.23)	Eye Open [POST] – μ (-0.23)	SAPS-I – First (0.29)	Urine – $\sum$ (-0.28)
Lactate – Min (0.19)	Urine [POST] – $\sum$ (-0.23)	Hep. Failure (0.20)	Eye Open – μ (-0.27)
Hep. SOFA – Min (0.19)	Shock Index [POST] – μ (0.21)	APS [D2] – Max (0.18)	APS – Max (0.25)
WBC – Min (0.19)	BUN/Creat. [OVER] – Min (0.17)	Cirrhosis (0.16)	BUN/Creat. – Min (0.22)
pH [MV] (-0.17)	SOFA [PRE] – μ (0.17)	Elective admission (-0.16)	Elective admission (-0.21)
Resp. Rate – μ (0.17)	Temperature [POST] – μ (-0.16)	Age [MV] (0.13)	Cirrhosis (0.20)
Shock Index – μ (0.15)	Elective admission (-0.16)	Mech. Vent. (0.13)	Hep. Failure (0.20)
CO – Max (-0.13)	Hep. Failure (0.16)	Ethnicity (0.13)	SOFA – μ (0.20)
Hem. SOFA – μ (0.13)	SOFA [POST] – Max (0.15)	Ringers [D1] (-0.13)	Temperature – μ (-0.17)
FiO2 – σ (-0.13)	Cirrhosis (0.15)	Glucose [D2] – μ (0.12)	Age [MV] (0.17)
BUN – Min (0.12)	Age [MV] (0.13)	Surgical ICU (-0.12)	Sedatives (-0.15)
Ringers (-0.12)	Pseudo. Monas. Aer. [OVER] (0.10)	Depression (-0.11)	Pseudo. Monas. Aer. (0.13)
Urine – Min (-0.12)	Glucose [PRE] – σ (0.09)	$PaCo_2$ Δ (-0.11)	SpO_2 [MV] (0.12)
Intub. (0.11)	Ringers [PRE] (-0.08)	Enterococcus SP [D2] (0.10)	Lactate – Min (0.12)
NeuroSOFA – Min (0.10)	MAP [POST] – Min (-0.07)	Eye Open [D2] – μ (-0.10)	Eye Open – Min (-0.11)
INR – Max (0.10)	Ethnicity (0.07)	HCO_3 Δ (-0.10)	WBC – Min (0.11)
DP – μ (0.10)	Lymphoma (0.07)	Temperature [D1] – μ (-0.10)	CO – Max (-0.11)
SpO_2 – μ (-0.10)	Mech. Vent. (0.07)	Temperature [D2] – μ (-0.10)	HE Length – Min (0.09)
Na – Min (0.09)	Motor Response [OVER] – Max (-0.06)	Bilirubin Δ (0.10)	Surgical ICU (-0.09)
Glucose – μ (0.09)	Glucose [PER] – Δ (0.06)	CVP [D2] [MV] (0.09)	Lymphoma (0.09)
Hem. SOFA – Min (0.09)	SpO_2 [POST] – μ (-0.06)	Lactate [D1] – Min (0.09)	Ethnicity (0.09)
INR – μ (0.08)	Surgical ICU (-0.06)	Urine Δ [MV] (0.08)	Depression (-0.09)
Sedatives – $\sum$ (-0.08)	SOFA [POST] – μ (0.06)	Bilirubin [D2] – Max (0.08)	pH [OVER] – σ (0.09)
$PaCO_2$ [MV] (-0.07)	Depression (-0.06)	SOFA Δ (0.08)	BUN/Creat. [OVER] – σ (-0.09)
Blood infection (-0.07)	Pressors [PER] – Δ [MV] (-0.06)	Lymphoma (0.08)	Mech. Vent. (0.08)
Platelets – σ (0.07)	FiO_2 [POST] – Min (0.06)	pH [D1] – σ (0.08)	Motor Response [OVER] – Max (-0.08)
Chloride – μ (-0.07)	NI ABP_{dia} [OVER] [MV] (0.06)	Urine [D1] – $\sum$ (-0.08)	Hep. SOFA [OVER] – σ (-0.08)

as $HE - OVER$ may indicates the presence of additional predictive factors. The next section therefore studies the relation between the temporal proximity to the outcome and the model performance in a more controlled setting.

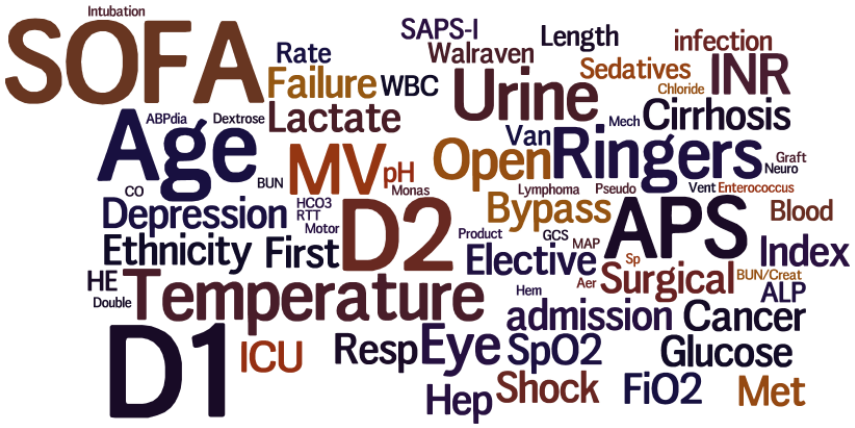


Figure 7.2: *Aggregated representation of variable importance. The size of each variable name on this "word cloud" is the number of occurrence of each word listed on table 7.3.*

7.2 Relation between model accuracy and proximity to the outcome

The work presented in this chapter uses data extracted from time windows (see Figure 7.1) that go beyond the traditional first 24 hours following admission to the ICU. In this context, it is of particular interest to assess how much of the observed improvement can be imputed to the temporal proximity with the outcome. In fact, it is generally well accepted amongst clinicians that prediction of a clinical event improves the closer one is to the event. Indeed, most clinical events are defined with simple rules and thresholds on physiological variables that can directly be captured from the data. Consequently, the closer in time to the event, the more accurate its prediction bound to be. Such a statement follows common sense and seems difficult to challenge.

However, reviewing works that looked at the prediction of hypotensive episodes in the MIMIC-II database present evidence that this may not always be the case. The problem is presented in figure 7.3, which represents a hypotensive episode with a predictive

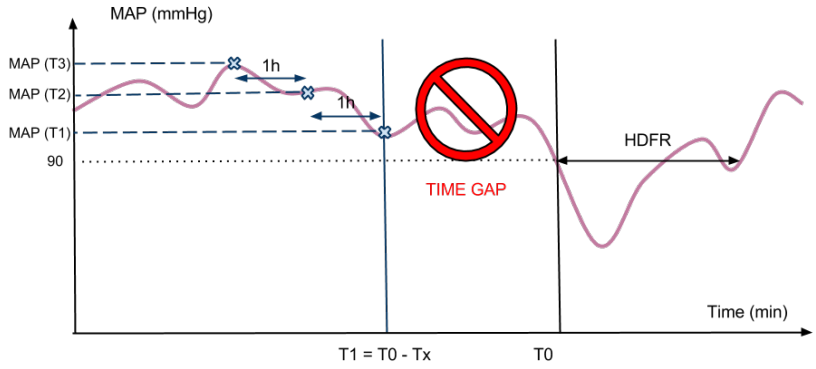


Figure 7.3: *This figure shows the Hypotension Despite Fluid Resuscitation (HDFR) as defined in Shavdia [255], the time gap considered T_x (that can be 30, 90, 120, 180 or 240 minutes) and the different values extracted from each piece of physiological data (only ABP_{Sys} plotted here): values at T1, T2 and T3 as well as the differences between T1-T2 and T3-T2*

window. First, in a population of septic patients, Shavdia [255] found that predictions made 120 minutes ahead of the hypotensive episodes were more specific than predictions made only 30 minutes prior to the event ($Spe_{120min} = 96\%$ and $Spe_{30min} = 87\%$, respectively at roughly equivalent AUROCs of around 95%). Second, on a more general ICU population Ghassemi [104] reported similarly counter-intuitive results and reasoned as follows:

- the time stamping accuracy near adverse events is degraded due to staff workload;
- there exists a specific pattern occurring two hours before the hypotensive episode allowing for better prediction at this time;
- patients' physiology is less stable shortly before the hypotensive episodes, potentially leading to incoherent predictions;
- staff intervention increases while patients reaches the threshold of hypotension, increasing the level of noise in the data.

Such a finding however is rare. On the contrary, Lee and Mark [175] worked on a very similar problem and used additional features including wavelet coefficients derived from vital signs. Their results did not corroborate the findings described above and model performance was, as expected, decreasing monotonically from the onset of the

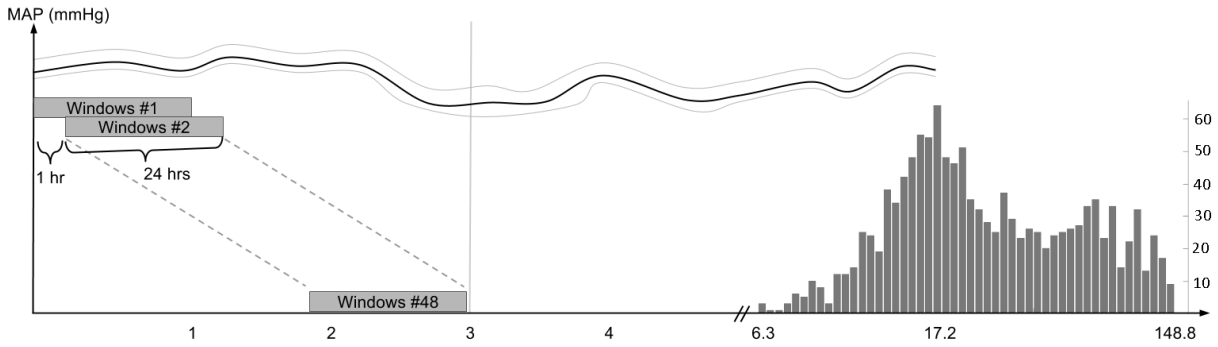


Figure 7.4: *(LEFT)Different time-window used for building estimates of severity over the first three days following admission to the ICU. (RIGHT) On the same time scale (days) is displayed the histogram of time of in-hospital death. For the sake of convenience the x-axis is log transformed from the first event indicated by a double slash sign (6.3 days). The y-axis on the right-hand side of the figure indicate the number of patients counted in each bin.*

hypotensive episode. Last but not least, these results obtained during the prediction of the onset of a hypotensive episode may not necessarily replicate on a dataset aiming at the prediction of mortality. In section 7.2.1, the relation between the temporal proximity to the outcome (in-hospital mortality) and the performance is therefore explored.

7.2.1 Comparison of models temporally closer to death

In-hospital mortality was predicted with all available physiological variables ($k = 220$) over a 24-hour time window. This time window was moved hour by hour from admission time ($t_0 = 0$) to two days after admission ($t_0 = 48$). All patients dying within the first 72 hours after admission ($n = 152$) were discarded from this study leaving $n = 2003$ patients for the analysis including $n = 469$ non-survivors (23.4%). The model used to make the predictions was the LASSO. In order to estimate the out-of-sample performance at each prediction time, a $k_f = 3$-fold cross-validation procedure was implemented in this section, using the very same fold as in previous chapters. In total there were $m = 3 \times 49 = 147$ models $M^t_{k_f}$ trained for which nearly three hundred thousand individual predictions were made.

The first experiment simply compares the AUROCs at different times t_0 after admission. The next two experiments concentrate on non-survivors for whom the interval

between the prediction time and the event (death) can be computed (since there is an occurrence of the event, unlike for the survivors). In the second experiment, the non-survivors are split into two groups of equal size according to the time separating the prediction from death. At each prediction time described in figure 7.4, the operating point is identified from the complete validation set (including survivors) and applied to non-survivors falling into the same fold to compute specificity. In the last experiment the relation between the residuals ($\epsilon_i = y_i - \hat{y}_i$) and the time interval between prediction and death is studied: for each non-survivor, $k = 48$ predictions made at different times were analysed. For each experiment, the performance metric (AUROC, specificity, and residuals for experiments 1, 2, and 3 respectively) is used as a regression variable within a linear multivariate model including time from prediction to the adverse event, rate of missing data, and fold number (coded as binary dummy variable).

7.2.2 Results

A statistically significant relation between time and AUROC, specificity, and the residuals was found. For the first study, figure 7.5 shows the regression line of AUROC against time that was associated with a statistically significant coefficient ($p < 0.001$). Figure 7.6 shows the regression analysis between time and specificity for the two groups (defined by the median time to death) with different intervals between prediction time and death, which varies from 6.3 to 17.2 days in the first group ($n = 234$) and from 17.3 to 148.8 days in the second ($n = 234$). The group that is most distant in time from the event does not show a statistically significant temporal coefficient ($p = 0.66$) in the regression, while the coefficient was found to be highly significant ($p < 0.001$) in the group including non-survivors that were closer in time to death.

Figure 7.7 plots the evolution of residuals in non-survivors against the interval between modelling time to predicted event (death). The residuals were used as a dependent variable in a linear multivariate analysis using the rate of missing data (respectively τ_{tr} and τ_t in the training and validation sets used to generate the prediction) and the

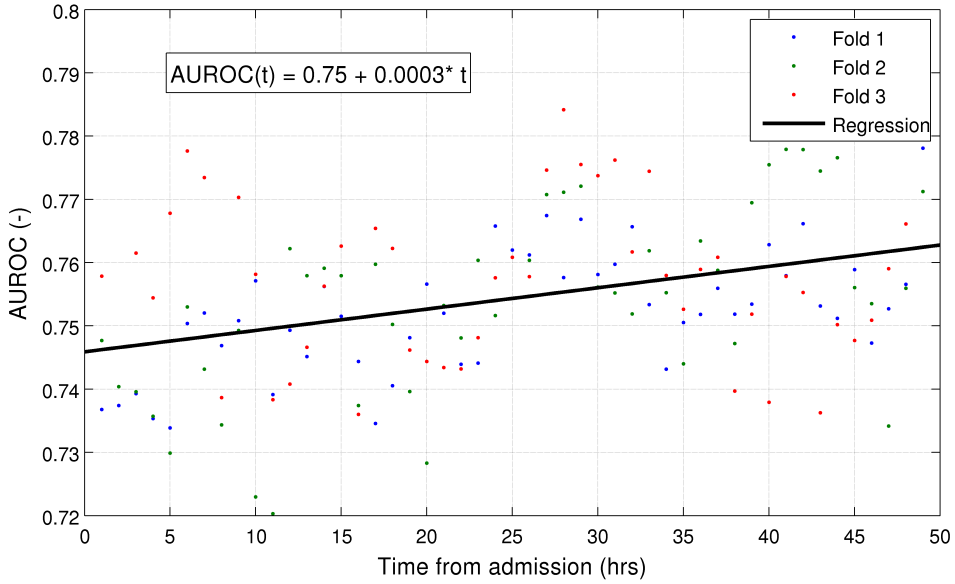


Figure 7.5: *Evolution of AUROC for models using data in 24-hour time window starting at different times after admission (x-axis). The plot show the results on three validation sets of size $n = 671$ (blue), 655 (green), and 665 (red). A linear regression is fitted to the data (black line) and the p-value associated with the time coefficient is statistically significant ($Prob(\beta/\sigma) = p < 0.001$) where β and σ denote the parameter value and its variance, respectively.*

time to death t_d as independent variables. The final relation used was

$$\epsilon_i = 0.23 - 0.014t_d + 0.016\tau_{tr} - 0.02\tau_t \tag{7.3}$$

and the respective p-values ($Prob(\beta/\sigma)$) were $p_t < 0.001$, $p_{\tau_t r} = 0.09$, and $p_{\tau_t} = 0.01$. The rate of missing data was consistent across prediction time and was found to be $0.23\ (0.18 - 0.29)$ for both the training and validation sets. However, only the rate of missing data in the validation set τ_t was found to be statistically significantly related to the residuals, while only a trend could be observed on the training set τ_{tr}.

7.3 Discussion

In this study, the focus on dynamic variables introduced a novel approach to the development of a hospital mortality prediction algorithm. In the first study, the dynamic information was extracted from physiological trends identified over the first two days

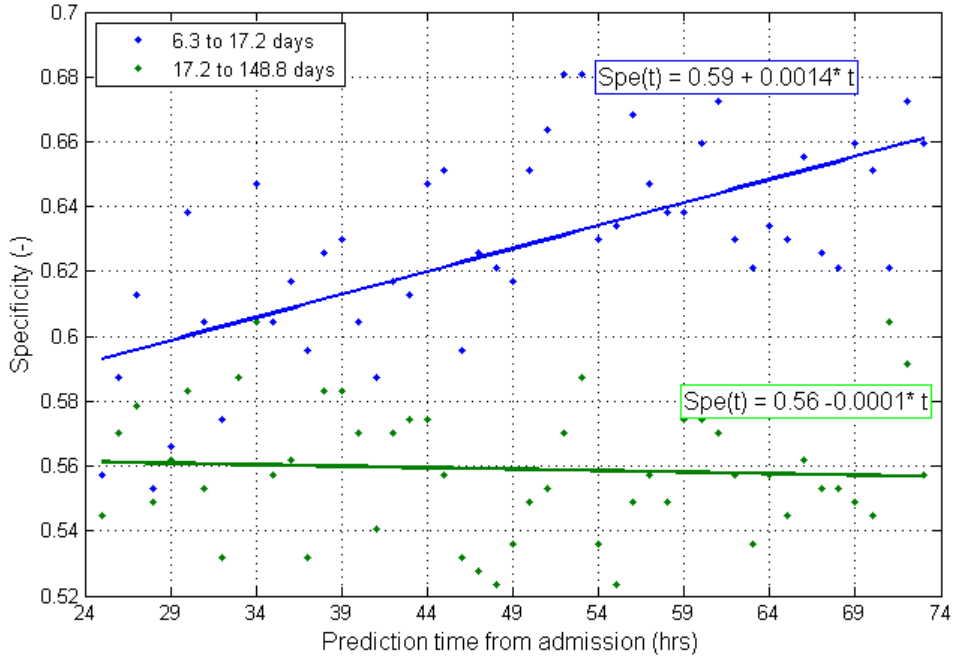


Figure 7.6: *Evolution of specificity (ability to identify non-survivors) on validation set for models using data for 24-hour time window starting at different times after admission (x-axis). The non-survivors were split into two groups of equivalent size with respect to the time separating the time of admission from the adverse event (death): from 6 to 17.2 days (blue, $n = 234$) and from 17.2 to 148.8 days (green, $n = 234$). Regression lines are fitted to both groups showing a statistically significant relation between the specificity and time only for the first group ($p < 0.001$ and $p = 0.66$).*

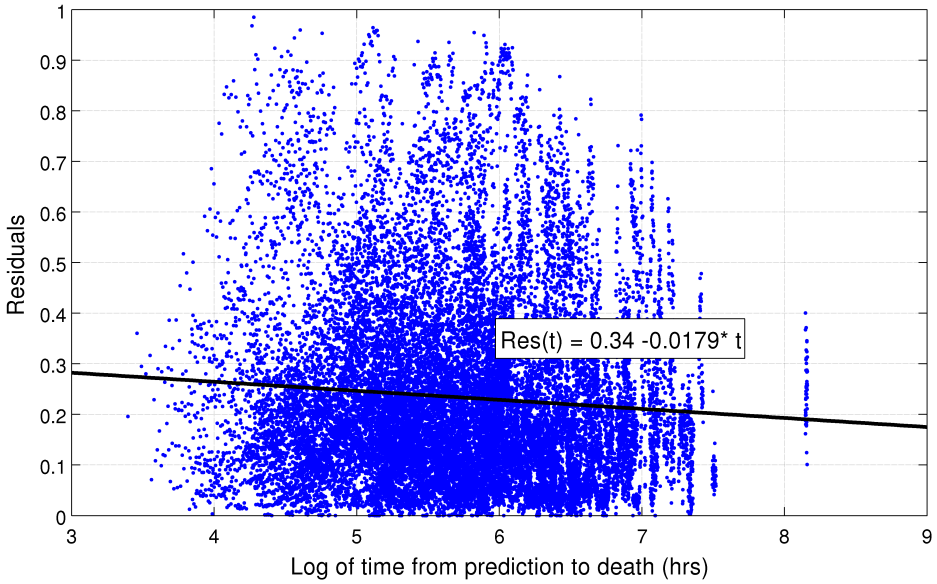


Figure 7.7: *Residual values for non-survivors ($n = 469$) in validation sets in relation to the interval between prediction time and death (x-axis). The predictions were made at different times after admission for each individual so that each patient is represented by $k = 48$ different data points separated by one hour. The regression line shows a statistically significant relation between residuals and time to the prediction ($p < 0.001$).*

following admission to the ICU. Then, it was hypothesized that patient response to both endogenous (hypotension) and exogenous (treatment) impulses may describe health status better than static information. For this reason, the second study focused on data surrounding a hypotensive episode, during which both hypotension and treatment were present.

Current prognostic scoring systems often predict similar outcomes for patients with the same comorbidities, severity of physiologic injury, and degree of organ dysfunction. In clinical practice, there is often wide inter-individual variability in outcome even when subjects fall within the same risk strata according to ICU admission scoring systems such as APACHE-IV. This may be because an important predictor of outcome, the individual's physiologic reserve [30], is not captured in these scoring systems. Physiologic reserve may account for the difference in clinical outcome for two patients with identical mortality risks (as traditionally defined by age, severity of illness and comorbidities) and treatment may have. Bion [30] places a large emphasis on the importance of cel-

lular processes in response to stress and oxygen delivery as the major determinant of this physiologic reserve, which is thought to vary between patients because of genetic differences.

Prior studies have attempted to measure aspects of the physiologic reserve. For example, Vallet et al. [278] demonstrated that in a uniform population of patients with sepsis and normal lactate levels, survivors had an increase in oxygen delivery in response to dobutamine; this finding was subsequently validated by Rhodes et al. [239]. Identification of subjects with relative adrenal insufficiency with the corticotropin stimulation test may capture another aspect of physiologic reserve [14]. The latter is likely to be dependent on the complex interplay between an individual's genetic background [133] and the physiologic insult. It is possible that after controlling for comorbidities, severity of insult and treatment, the dynamic variables surrounding a hypotensive event may allow us to determine the contribution of an individual's physiologic reserve to prognosis, thus allowing better individual (as opposed to group) predictions of hospital mortality in patients with septic shock.

The different models presented in this section strengthen the importance of some admission variables that were identified in previous chapters: age, elective admission, bypass surgery, as well as some CHC (metastatic cancer, hepatic failure, depression, and cirrhosis). During the process of de-identification, age was removed for any patients older than 90 years old, which explains why $Age\ [MV]$ is so closely related to severity. Again, depression was found to be associated with an overall lower severity. Potential confounders such as treatments for chronic depression Serotonin-norepinephrine reuptake inhibitor (SNRI) and Selective serotonin reuptake inhibitors (SSRI) were extracted from the patients' discharge summary and included in the study without changing the results. Other physiological variables also confirmed their relation to severity: cumulative sum of urine, temperature, lactate, SpO_2, glucose, and FiO_2, which altogether seem very consistent with the population studied. In addition to this, the presence of two microorganisms was also found to be associated with an increased severity: $Enteroccocus\ Sp.$

and $Pseudo.\ Monas.\ Aerus$, that were previously related to worst outcome in such a population [6, 60, 225]. Interestingly, the presence of blood infection over the first two days was detrimental, while its identification over the hypotensive episode seemed beneficial.

Severity scores used for benchmarking are restricted to data collected in the first twenty four hours in an attempt to not incorporate local practice and therefore best reflect patient physiological condition at admission. By definition then, these scores do not incorporate any information related to treatments or interventions. However, we have seen in section 5.3.3 that the inclusion of such co-variates over the first twenty-four hours improved the score accuracy. This result clearly indicate that treatment actions during the first day in the ICU influences prognosis. Because the patient physiology is strongly influenced by external intervention, it does not seem reasonable to assume that the worst value of a physiological covariate over the first day is independent of local practice. For instance, a lower diastolic blood pressure reading over a day tells very little without information related to the fluid balance and the potential use of vasopressor medication. By definition, the study presented here incorporates information over time-periods where a patient's physiology is strongly influenced by therapy and it seemed necessary to account for it by including variables reflecting interventions. Interestingly, not all treatment were found to be associated with a worst outcome: for instance, patients undergoing bypass surgery, or who were administered ringers or dextrose had a lower severity. Conversely, sedation, mechanical ventilation and intubation were as expected associated with a higher risk of dying.

The incorporation of dynamic information into the models was three-fold: (i) the inclusion of a Δ variable as defined by equations 7.1 and 7.2, (ii) the selection of variable standard deviation (σ), and (iii) the selection of the same variable extracted at different times. For instance, in the first study (Model $D1D2-\Delta_{24}$), the rise in bicarbonates (HCO_3 [Δ_{24}]) between day one and two seemed beneficial, and the increase in volume of Ringers saline solution administered (Ringers [Δ_{24}]) clearly indicated a worst outcome. Then, in (ii) the standard deviation of pH over the first day (pH [D1] – σ) was found to be related

to severity and finally, (iii) the mean temperature was incorporated into the model over day 1 and day 2. These findings only report the most important variables identified by LASSO and displayed in table 7.3 and illustrate the mechanisms at stake for the $k \approx 70$ variables left in the model. Dynamic components clearly account for the statistically significant gain in performance observed, while the data collected to build Model $D1D2$ were strictly collected over the same period.

During the second study (Model $HE - \Delta_{24}$), dynamic components were equally represented within the most relevant features identified by LASSO. In particular, the onset of RRT immediately after the first hypotensive episode (RTT [Δ_{HE}]) possibly identifies renal failure that is subsequent to lack of perfusion, which may thereby adequately reflect patients' physiologic reserve. Similarly, the variation of glucose during the day preceding the onset of the hypotensive episode (Glucose [PRE] – σ) was associated with a worst outcome, which possibly reflects metabolic pathways and perhaps indirectly physiologic reserve. Likewise, SOFA was selected before and after the event showing the independent contribution of this variable taken around the hypotensive episode. The fact that the addition of these dynamic components did not translate into a statistically significant improvement in predictive performance certainly has many origins. Amongst them, the rise in performance induced by the temporal proximity with the outcome (demonstrated in this chapter) has possibly hindered the identification of an effect that could be statistically significant on a population of such a size.

Conclusion

The objective of this chapter was to estimate the potential benefit of a shift in approach toward the estimation of patient's severity in the ICU from static models (at a point in time) to *dynamic* models (looking at fluctuations in the physiology). Firstly, the evolution of physiological variables over time was investigated. Then, the evolution of a patient's physiology over specific internal (hypotension) and external (treatment) events was ex-

amined.

Results suggest that incorporating the evolution of some variables rather than their absolute value is beneficial to model performance. This benefit was clearly demonstrated by the increased discriminatory power of models $D1 - \Delta_{12}$ and $D1D2 - \Delta_{24}$ over $D1$ and $D1D2$, respectively. Models incorporating data surrounding the hypotensive episode that were much closer in time to the outcome (i.e. death) all presented increased performance. In that context the potential benefit of the *dynamic* approach could not be demonstrated in a way with statistical significance. Finally, the relation between temporal proximity to the outcome and model performance was illustrated in different studies all confirming the presence of such a mechanism.

Life can be defined as "a process or system characterized by constant change, activity or progress", which is the exact definition of *dynamic*. Paradoxically, homoeostasis – the process by which system constants such as temperature are preserved – is itself a dynamic phenomenon because it is the result of numerous positive and negative regulation loops associated with different time constants. This non-explicit dynamic nature of homoeostasis has certainly contributed to the fact that early approaches toward the estimation of life functions (and inversely the severity of a disease) were initially based on static data analysis. Another plausible cause lies in the data collection burden that dramatically constrained (and still does to some extend) the progress of medical science. Thankfully, the MIMIC-II database unlocked this barrier and finally allowed the in-depth exploration of such hypothesis, revealing the potential of this approach.

Chapter 8

Conclusion

8.1 Thesis overview

In chapter 1 we introduced sepsis as the second largest killer in ICU after coronary disease (it was associated with an estimated cost of $\$17b$ every year in the US). We then presented what is currently known about the underlying mechanisms of sepsis revealing a complex entanglement of physiological responses triggered by the presence of external agents and leading to several degrees of physiological dysfunction: alteration of micro-circulation, haemodynamic shock, and multiple organ failure. Despite this knowledge, sepsis and its complications remain only partly understood. In fact, what is known about sepsis has evolved over past decades, translating into significant changes in its definition.

The aim of this work was to improve the estimation of severity in a population of patients with sepsis and hypotension. This was done with two main objectives: (1) gain understanding of the underlying pathophysiological processes and (2) raise the overall performance in estimating the risk of in-hospital mortality in order to make such tools usable at bedside. In order to do so, a cohort was identified from the MIMIC-II database (v2.26), which contains the data of more than thirty thousand patients admitted to the ICUs at BIDMC between 2001 and 2008. As described in chapter 2, the cohort was com-

posed of 2,143 adult patients with sepsis (according to the Angus et al. [6] definition), hypotension, and treatments for it (fluids and pressors).

Traditionally the estimation of a patient's severity in the ICU is achieved through mortality prediction models, which link a binomial outcome (often in-hospital mortality) to an ensemble of explanatory variables. APACHE and SAPS are two common approaches towards mortality prediction that exist in different versions as detailed in chapter 3: they include Chronic Health Condition (CHC), demographics, admission data and physiological status derived from the first 24 hours after admission to the ICU. The implementation of state-of-the-art models validated the superiority of the fourth version of APACHE ($Nll = 385.8$ and $AUROC = 71.15$%) while its performance on our cohort is found to be on the lower side of what was found in the literature. A first level of customization consisted in identifying new coefficients for the model that were derived from our cohort data and leading to a slight statistically significant improvement in discriminatory power ($AUROC = 73.3$%, $IDI = 2.3$, $p_{IDI} = 0.04$).

In order to improve on this performance, the use of advanced machine learning technique was suggested and implemented in chapter 4: SVM and RF were chosen for their non-linear properties and because they represent of state-of-the-art machine learning techniques. Consequently, the gain provided with respect to logistic regression, the method of choice in mortality prediction models, could be evaluated, only showing a slight improvement for RF that was not found to be statistically significant ($Nll = 371.2$, $AUROC = 74.4$%, all $p > 0.05$). In addition to this a shift in approach was suggested from expert-driven selection of variables to objective feature selection techniques: the LASSO and *ReliefF* algorithms were described and implemented. The LASSO offered the best performance with a statistically significant improvement against logistic regression ($Nll = 369.4$, $AUROC = 75.1$%, $IDI = 1.7$, $p_{IDI} = 0.05$). Finally, the top-two strategies presented at the latest scientific challenge aiming at mortality prediction on a similar dataset from the MIMIC-II database were also implemented and compared: the cascaded-SVM and the BEF. The BEF algorithm offered a greater improvement in

discriminatory power than LASSO but had an overall lower performance ($Nll = 374.4$, $AUROC = 76.16\%$, $IDI = 3.6$, $p_{IDI} = 0.01$).

The use of machine learning techniques demonstrated a benefit over the use of logistic regression on the subset of variables initially used to build APACHE-IV model. Yet, as presented in chapter 5, the added value of nearly five thousand variables available in the MIMIC-II database still remained unexplored. Furthermore, additional features that are derived from these variables might equally contribute to the model: for example, non-linear combinations of physiological variables including existing severity scores and binary-coded variables to indicate the presence of a missing value. Given the favourable balance between complexity, computational cost and performance, the LASSO algorithm was seen as the best tool to carry out this study. The addition of all available covariates (not restricting the study to those identified by the APACHE or SAPS groups) statistically significantly boosted the performance ($Nll = 348.7$, $AUROC = 78.9\%$, $IDI = 4.1$, $p_{IDI} < 0.01$). The addition of physiologically meaningful non-linear combinations of variables, existing severity scores, and binary-coded missing values all further improved the model performance independently as well as when used together ($k = 414$ variables, $Nll = 337.8$, $AUROC = 80.4\%$). Finally, to address the potential limitations of the LASSO framework, all the aforementioned machine learning techniques were applied to this new subset of variables showing at best equivalent performance.

Physiological processes are believed to be intrinsically non-linear and the superiority of the LASSO algorithm over more complex non-linear approaches remains unexplained, yet in line with the literature. One hypothesis suggested is that the two-step procedure for feature selection and model fitting is suboptimal. Instead a framework allowing simultaneous feature selection and parameter pruning was investigated and implemented with a GA in chapter 6. Different fitness functions to be optimized through the GA framework enabled the exploration of different hypotheses, which all eventually led to equivalent performance. The first fitness function was a simple log-likelihood function meant to select the most relevant subset of features in a way similar to that of the LASSO. It

achieved similar performance ($Nll = 341.8$ and $AUROC = 80.8$%), thereby validating the overall framework. The second GA implemented a fitness function inspired by ARD that was meant to embed the concept of Occam's razor and therefore automatically identify the optimal model size. This framework offered the best performance of all GAs ($k = 34$ variables, $Nll = 333.1$, $AUROC = 80.04$%) at a very attractive computational cost. Finally, the SVM-based fitness function did not achieve the best performance ($Nll = 337.9$, $AUROC = 80.0$%) but compared advantageously to previous occurrences of a similar SVM model. Yet, the marginal benefit presented by the GA framework on this dataset does not seem to justify the increase in complexity and computational cost.

The models presented in this work together with the studies surrounding their design led to the identification of previously reported markers of severity in sepsis, severe sepsis, and septic shock: age, metastatic cancer, chronic liver disease, acute kidney and lung dysfunction, lactate levels, WBC count, and more. Additional variables such as the time to the first hypotensive episode, the amount of insulin given, the use of sedatives, the ethnicity, and the presence of chronic depression for example may be the substrate for the generation of new hypotheses. In terms of clinical use, none of the models presented so far could reasonably be deployed at the bedside even though the accepted threshold of usability ($AUROC > 80$%) was met. In a last attempt to further improve model performance, a final change in paradigm was suggested: rather than looking at physiological data over a static time-window ("the first 24 hours"), the model should incorporate *dynamic* information. To capture this dynamic information two studies were carried out. In the first study, models derived with the LASSO algorithm from the first 24 and 48-hour time-windows were compared with dynamic models derived over the same period of time with respect to admission to the ICU. The dynamic models simply consisted of two consecutive time-windows between which a simple difference was computed and used as an independent feature. The two dynamic models offered very similar performance ($Nll = 321.8$ and $AUROC = 82.7$%) and were both found to be superior to the "static" model computed over the same time period with respect to admission ($IDI = 3.51$ and

$IDI = 2.46$ with both $p_{IDI} < 0.05$). In an attempt to capture physiological response to the hypotension and the treatment given for it, the second study was carried out in a similar way but with time-windows centred around the hypotensive episode. Because of temporal proximity with the outcome, the model developed with a single time-window spread over the hypotensive episode gave significantly better performance ($Nll = 304.5$ and $AUROC = 85.3$%). With respect to this baseline, the dynamic model could however not demonstrate a statistically significant improvement ($Nll = 299.2$, $AUROC = 85.6$%, $IDI = 0.2$, $p_{IDI} = 0.42$).

We have shown using this database that expert knowledge provides worse results than an objective selection of clinical covariates based on specific feature selection techniques [191]. However, experts have presumably been exposed to a much larger sample of patients, practices, services and hospitals. It is therefore reasonable to assume that they can identify a representative subset of variables more reliably than any algorithms derived from the data collected in one hospital. In other words, expert knowledge may reflect practice variation and case-mix, while the database simply cannot offer this. There is one way to answer this question: collect similar data from different centres and compare the two approaches. Whatever the results of this experiment, the data will then be available globally to capture the geographical and practice variation between institutions by means of the same objective techniques. In addition to this, the presence of international data may even cover domains of expertise rarely covered by a single clinician, since international exchanges are unfortunately too short or too rare. Similarly, the storage and analysis of historical data may capture cyclic or sporadic events (such as seasonal flu or epidemics) to which a young, or even more experienced practitioner, may not necessarily have been exposed. Finally, the small changes in incidence for different diseases may also be incorporated into such a model whereas clinicians may not immediately capture the increasing importance of a novel co-variate.

8.2 Future work

8.2.1 Feature extraction from waveforms

HR response to environment changes is mediated by the ANS that plays a central role in the mediation of the inflammatory response to infections as described in section 1.2.3. In particular, it has long been demonstrated that "short"- and "long"-term variabilities relate to the vagal and sympathetic function of the ANS, respectively. Increased variability has been associated with a faster recovery in a wide range of patients [47], including critically-ill patients and those with sepsis [98]. Similarly, decreased variability has been consistently associated with worse outcome during sepsis in adults [165, 21, 58], neonates [170] and animal models [90]. Most published work on the topic concentrate on Heart Rate Variability (HRV) analysis [267] that derives several features from HR time series in order to estimate relative activation in different frequency bands believed to be related to specific functions of the ANS. Because such metrics can theoretically be derived from any bedside ECG recording, they have been widely studied in a wide range of critically ill patients. Most recently, Riordan et al. [242] related the loss of heart rate complexity to mortality in a cohort of 2,178 trauma patients regardless of severity of injury. In this work, the heart rate complexity was estimated with Multi-Scale Entropy (MSE) [69], a technique from the field of information theory, over a period of six hours directly following admission to the ward. These results were corroborated in 12 haemorrhage pigs by Batchinsky et al. [23] who subsequently confirmed the results in 31 trauma patients showing consistency of findings with records as short as ten minutes long.

In addition to the clinical data that is stored in MIMIC-II, the Physionet platform also offers high-resolution physiological records in the MIMIC-II *waveform* database (version 3) [248]. This database must be distinguished from the *clinical* database (version 2.6) even though a partial match between records and ICU patients can be established so that about 10% of the patients present in the clinical database can be linked to one or more physiological waveform record. Each record contains one or more raw waveforms

that are directly extracted from the bedside monitors with a typical sampling frequency of 256Hz and a resolution of 12 bits although different combinations are also frequently found. A record is defined by its starting date (approximately matching the randomly shifted admission dates in the clinical database), its length and a few fields describing the signals recorded. Types of waveforms present in the database include: ECG, ABP, SpO_2, CVP, as well as other less common signals such as PAP, which were all introduced in section 1.2.2. Alternatively, records can contain a minute-by-minute time-series, such as heart rate derived from the raw ECG. These trends however contain processing that is specific to the bedside monitor (like filtering and artefact rejection) and it is usually assumed to be safer if the heart rate time-series is derived directly from the high-resolution waveforms [63].

Extracting high-temporal resolution data from the MIMIC-II waveform database requires a match to a specific ICU: the waveform database provides each record with a subject ID and a time-stamp indicating the beginning of the record. This time-stamp is assumed to be readjusted to the admission date that was randomly shifted during the process of de-identification. Each record can therefore be related to an ICU stay after the examination of admission and discharge dates and times. Yet, this process can be complicated by several factors. Firstly, monitor clocks are rarely synchronized exactly with the HIS and delays can reach up to an hour during energy saving times. In addition to this, it is not infrequent to have the bedside monitor fitted to a patient in the operation theatre or in a different service potentially hours before the patient's admission to the ICU. From the $4,897$ waveform records available in the current version of the waveform database (v3.1), $4,762$ records (97.2%) started within a day of an ICU stay. From these, 321 records (6.7%) started before admission with an average difference of 4.4 hours. Finally, mapping the subset of patients described in section 2.2.1 with the matched subset of records leads to the identification of 218 waveforms with an average length of 2.5 $(1.0-5.1)$ days. Similarly, 239 numeric records with a length greater than five hours could be matched to our population from $5,2666$ records (containing HR trends extracted from the bedside

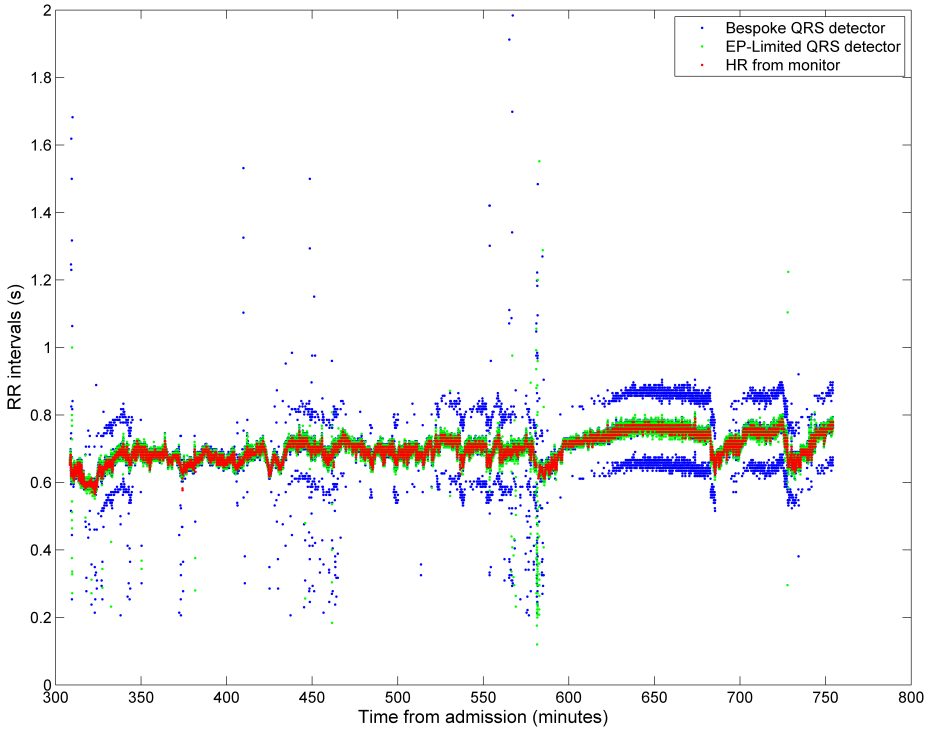


Figure 8.1: *Tachogram for patient 2064 showing different HR estimations: State-of-the-art (green), bespoke QRS detector (red) and that extracted from the bedside monitor (red). Episodes of arrhythmia are characterized by two extra bands at about 100ms below and above the average RR line. The lower bands reflects premature beats (probably, Premature Ventricular Contraction or PVC), while the upper band is associated with a subsequent compensatory pause. These events are missed by the EP-Limited detector (green) and possibly filtered out by the bedside monitor algorithms (red).*

monitor) with an average length of $2.8\ (1.1 - 5.3)$ days. The population of patients who had at least one matched record had worst outcomes (in-hospital mortality of 28.3%) than the general population and they all had strictly positive SOFA scores. This finding is consistent with the fact that a population invasively monitored is usually more severely ill.

The first step before deriving features from ECG waveforms is to adequately identify the R peaks; a process largely complicated in the ICU by movement artefact as well as physiological derangement. For instance figure 1.1 shows an example where the T-wave is larger in amplitude than the R-wave, which is unusual. Figure 8.1 shows the comparison of three different QRS detectors: the first one (red) is the second-by-second

trend generated by the monitor; the second (green) is an open-source QRS detector [1] that was initially based on the detector provided with the Waveform Database (WFDB) Physionet toolbox [199, 107]; the third (blue) is a bespoke QRS detector inspired by Pan and Tompkins [217], Hamilton [117] and subsequently modified to benefit from the offline nature of the analysis. Different Signal Quality Indices (SQI) can also be derived from the high-resolution ECG time-series. HRV [267] and complexity (MSE) analysis can then translate these signals (RR and SQI) into features in order to investigate their relation to patients' severity.

Early results on this data suggest that complexity measures derived from ECG in the ICU do contain discriminatory information about patient severity [192]. Features derived from the HRV analysis seemed to offer greater predictive power than those obtained with MSE. In both cases however, these complexity metrics used in combination with an existing severity score (APACHE-IV) showed an improvement that was not found to be statistically significant. Interestingly, complexity measures derived from the SQI did also show discriminatory power, possibly reflecting the level of intervention or sedation. Such work and similar approaches [208] should be continued in order to develop the predictive potential of the most commonly available non-invasive monitoring modality. In order to allow this, modern data collection systems must be developed so that large amount of high-resolution data can be routinely collected and linked to clinical information.

8.2.2 Towards the digital ICU

ICU spending represents an estimated 1% of GDP in developed countries [219, 115, 116] steadily increasing as a result of population ageing. In the context of growing economic stresses, the pressure on healthcare is such that delivery of critical care must be rationalized. A fact well recognized by seasoned intensivists is the information gap that exists in critical care practice, i.e. the absence of evidence to guide most decisions, which is ironic given the "data overload" in the ICU. The intensivist often comes out empty-handed from

[1] EP-Limited available at www.eplimited.com

a search for a relevant prospective randomized controlled trial (PRCT), the gold standard in evidence-based medicine, although fewer than 15% of them end with a clear benefit [214]. In addition to this, an estimated 70% of interventions in the ICU are based on marginal evidence at best [288]. The paradox characterized by the data-overload and the lack of evidence is clearly exacerbated when it comes to making real-time bedside decisions. Tools now exist to direct this real-time deluge of data to a cloud infrastructure in order to facilitate decision making at bedside.

Few attempts have been made to adequately exploit the data produced during patients' stay in the ICU. General severity scores (APACHE, SAPS) [313, 201, 204] represent such attempts. Additionally, companies like Cerner [167] routinely collect equivalent data on a daily basis all around the world. These attempts however still lack the sort of temporal granularity necessary to capture individual variation in the response to a physiologic insult for improved risk-benefit calculation and outcome prediction. In fact, the MIMIC-II database is unique in capturing such highly granular ICU data [248] and we hope that this work has demonstrated its potential in tackling the aforementioned paradox. Unfortunately MIMIC-II only captures about 40,000 unique patients from a single centre.

The Paris hospital trust (AP-HP) has initiated the development of an integrated information system across 37 hospitals that cover 1.2 million admissions per annum. This system will soon encompass more than fifty ICUs. The implemented AP-HP Electronic Healthcare Record (EHR) handles high-granularity data including minute-by-minute changes in heamodynamics, waveforms and other physiologic signals as well as time-stamped treatments and their dosage, e.g. fluids, blood products, medications, procedures. If the AP-HP project is noticeable by its size and level of progress, similar projects are mushrooming all over the world and in particular in Europe. Together with emerging technologies for storage and analysis of "Big Data", there is a unique opportunity to change the paradigm of medical knowledge acquisition and revolutionize clinical practice in the ICU.

The intensive collection of ICU data will allow the transfer of clinical events to computer science researchers paving the way for the validation of clinical concepts. Additionally, the outcome of research on retrospective data will undoubtedly raise worthwhile hypotheses ahead of costly and potentially harmful human trials [129]. This is not an argument for replacing the existing paradigm for evidence generation (RCTs) but rather to support and complement it.

The exploitation of such an infrastructure will require multidisciplinary research and the development of innovative solutions in the fields of massive data integration, knowledge extraction, decision support, intensive computation and simulation in order to achieve a learning health system and a system that continuously uses health care data to answer important questions that matter to patients and their health care providers.

The objective of such a project is therefore to develop and deploy a health information technology platform fuelled with high-resolution ICU data in order to generate evidence and drive the delivery of optimal care at the bedside in ICUs. In particular, it could target the following applications:

- Pharmacovigilance with existing unsupervised algorithms for novelty-detection;
- Early warning system for hypoperfusion, infection and cardiac events [303, 304];
- Early stopping system for mechanical ventilation, sedation, vasopressor therapy, and antibiotics [34];
- Intervention benefit estimator for red blood cells transfusion and blood sample withdrawals [61];
- Drug dosage recommendation capturing heterogeneity of treatment effect [166].

To summarize, massive amounts of clinical data will soon be available for real-time analysis thanks to efforts made at leading clinical institutions. Novel technological developments in the field of big data allow the storage and analysis of such data in real-time, paving the way for revolutionary applications. Given the cost of our critical care system

(1% of GDP and steadily increasing) as well as the lack of rationality that characterizes its delivery (1 in 7 positive RCT, 70% of interventions relying on weak evidence), the collection, de-identification and dissemination of this data offers a promising approach for the next medical revolution in critical care.

8.3 Conclusion

To conclude, starting from the traditional severity scores, different levels of customization were implemented : calibration coefficients, model coefficients, and an initial subset of features. After the customization of existing approaches, which proved beneficial, new modelling techniques were introduced, the extraction of data collected after the first 24 hours and the incorporation of dynamic information. Interestingly, the gain in performance did not necessarily come from the use of more sophisticated techniques. In fact the most important model improvement came from (1) the use of additional covariates and their transform and (2) the use of data subsequent to ICU admission. That being said, the use of sophisticated algorithms combined with a data-driven approach to feature selection was found to be a viable approach to outcome modelling in patients with sepsis and hypotension. While further studies on additional ICU populations are needed to validate this approach and these findings, this study demonstrates that such an approach has the potential to provide better predictions of hospital mortality, highlighting the role that clinical data mining will increasingly play in both knowledge generation and the way medicine is practised.

Bibliography

[1] Swapna Abhyankar, Kira Leishear, Fiona M Callaghan, Dina Demner-Fushman, Clement J McDonald, et al. Lower short-and long-term mortality associated with overweight and obesity in a large cohort study of adult intensive care unit patients. *Critical Care*, 16(6): R235, 2012.

[2] Jerome Aboab, Veronique Sebille, Mercé Jourdain, Jacques Mangalaboyi, Miloud Gharbi, Arnaud Mansart, and Djillali Annane. Effects of esmolol on systemic and pulmonary hemodynamics and on oxygenation in pigs with hypodynamic endotoxin shock. *Intensive care medicine*, 37(8):1344–1351, 2011.

[3] Horacio J Adrogué and Nicolaos E Madias. Hypernatremia. *New England Journal of Medicine*, 342(20):1493–1499, 2000.

[4] Bekele Afessa, Mark T Keegan, Rolf D Hubmayr, James M Naessens, Ognjen Gajic, Kirsten Hall Long, and Steve G Peters. Evaluating the performance of an institution using an intensive care unit benchmark. *Mayo Clin Proc.*, 80(2):174–180, 2005.

[5] Douglas G Altman, Yvonne Vergouwe, Patrick Royston, and Karel GM Moons. Prognosis and prognostic research: validating a prognostic model. *BMJ: British Medical Journal*, 338, 2009.

[6] D C Angus, W T Linde-Zwirble, J Lidicker, G Clermont, J Carcillo, and M R Pinsky. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. *Critical Care Medicine*, 29(7):1303–1310, Jul 2001.

[7] Derek C. Angus. Understanding the lingering consequences of what we treat and what we do. *Critical Care*, 8(2):103–104, Apr 2004. doi: 10.1186/cc2838. URL `http://dx.doi.org/10.1186/cc2838`.

[8] Derek C. Angus. The lingering consequences of sepsis: a hidden public health disaster? *JAMA*, 304(16):1833–1834, Oct 2010. doi: 10.1001/jama.2010.1546. URL `http://dx.doi.org/10.1001/jama.2010.1546`.

[9] D. Annane. Septic shock. *The Lancet*, 365:63–78, 2005.

[10] D Annane, E Bellissant, V Sebille, O Lesieur, B Mathieu, JC Raphael, P Gajdos, et al. Impaired pressor sensitivity to noradrenaline in septic shock patients with and without impaired adrenal function reserve. *British journal of clinical pharmacology*, 46:589–597, 1998.

[11] D. Annane, V. Sebille, G. Troche, J. C. Raphael, P. Gajdos, and E. Bellissant. A 3-level prognostic classification in septic shock based on cortisol levels and cortisol response to corticotropin. *JAMA*, 283(8):1038, 2000.

[12] D. Annane, P. Aegerter, M. C. Jars-Guincestre, and B. Guidet. Current epidemiology of septic shock: the CUB-Rea network. *American Journal of Respiratory and Critical Care Medicine*, 168(2):165, 2003.

[13] Djillali Annane. Corticosteroids for severe sepsis: an evidence-based guide for physicians. *Annals of intensive care*, 1(1):1–7, 2011.

[14] Djillali Annane, Véronique Sébille, Claire Charpentier, Pierre-Edouard Bollaert, Bruno François, Jean-Michel Korach, Gilles Capellier, Yves Cohen, Elie Azoulay, Gilles Troché, Philippe Chaumet-Riffaud, Philippe Chaumet-Riffaut, and Eric Bellissant. Effect of treatment with low doses of hydrocortisone and fludrocortisone on mortality in patients with septic shock. *JAMA*, 288(7):862–871, Aug 2002.

[15] Virginia Apgar. A proposal for a new method of evaluation of the newborn infant. *Anesthesia & Analgesia*, 32(4):260–267, 1953.

[16] Y. Arabi, N. Al Shirawi, Z. Memish, S. Venkatesh, and A. Al-Shimemeri. Assessment of six mortality prediction models in patients admitted with severe sepsis and septic shock to the intensive care unit: a prospective cohort study. *Critical care*, 7(5):R116–R122, 2003.

[17] F. Arnalich, J. López, R. Codoceo, M. Jiménez, R. Madero, and C. Montiel. Relationship of plasma leptin to plasma cytokines and human survival in sepsis and septic shock. *Journal of Infectious Diseases*, 180(3):908–911, 1999.

[18] L. Arregui, D. G. Moyes, J. Lipman, and L. Fatti. Comparison of disease severity scoring systems in septic shock. *Critical Care Medicine*, 19(9):1165, 1991.

[19] A. Artero, R. Zaragoza, J. J. Camarena, S. Sancho, R. Gonzalez, and J. M. Nogueira. Prognostic factors of mortality in patients with community-acquired bloodstream infection with severe sepsis and septic shock. *Journal of Critical Care*, 25(2):276–281, 2010.

[20] Gavin Barlow, Dilip Nathwani, and Peter Davey. The CURB65 pneumonia severity score outperforms generic sepsis and early warning scores in predicting mortality in community-acquired pneumonia. *Thorax*, 62(3):253–259, 2007.

[21] Douglas Barnaby, Kevin Ferrick, Daniel T Kaplan, Sachin Shah, Polly Bijur, and E John Gallagher. Heart rate variability in emergency department patients with sepsis. *Academic emergency medicine*, 9(7):661–670, 2002.

[22] Steven L Barriere and Stephen F Lowry. An overview of mortality risk prediction in sepsis. *Critical care medicine*, 23(2):376–393, 1995.

[23] Andriy I Batchinsky, James E Skinner, Corina Necsoiu, Bryan S Jordan, Daniel Weiss, and Leopoldo C Cancio. New measures of heart-rate complexity: effect of chest trauma and hemorrhage. *The Journal of Trauma and Acute Care Surgery*, 68(5):1178, 2010.

[24] J. D. Baumgartner, C. Bula, C. Vaney, M. E. I. M. Wu, P. Eggimann, and C. Perret. A novel score for predicting the mortality of septic shock patients. *Critical Care Medicine*, 20(7): 953, 1992.

[25] Michael Beebe, J Dalton, Martha Espronceda, and Desiree D Evans. *CPT 2007 professional edition (CPT/current procedural terminology [professional edition])*. Chicago, IL: American Medical Association Press, 2006.

[26] Rinaldo Bellomo, Claudio Ronco, John A Kellum, Ravindra L Mehta, Paul Palevsky, et al. Acute renal failure–definition, outcome measures, animal models, fluid therapy and information technology needs: the second international consensus conference of the acute dialysis quality initiative (ADQI) group. *Critical care*, 8(4):R204, 2004.

[27] Rodney P Bensley, Shunsuke Yoshida, Ruby C Lo, Margriet Fokkema, Allen D Hamdan, Mark C Wyers, Elliot L Chaikof, and Marc L Schermerhorn. Accuracy of administrative data versus clinical data to evaluate carotid endarterectomy and carotid stenting. *Journal of vascular surgery*, 2013.

[28] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13:281–305, 2012.

[29] F Bion, TC Aitchison, SA Edlin, and I McA Ledingham. Sickness scoring and response to treatment as predictors of outcome from critical illness. *Intensive care medicine*, 14(2): 167–172, 1988.

[30] J. F. Bion. Susceptibility to critical illness: reserve, response and therapy. *Intensive Care Medicine*, 26 Suppl 1:S57–S63, 2000.

[31] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer Sciences, 2006.

[32] Christopher C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[33] C.M. Bishop. Neural networks and their applications. *Review of scientific instruments*, 65(6): 1803–1832, 2009. ISSN 0034-6748.

[34] Bronagh Blackwood, Fiona Alderdice, Karen Burns, Chris Cardwell, Gavin Lavery, and Peter O'Halloran. Use of weaning protocols for reducing duration of mechanical ventilation in critically ill adult patients: Cochrane systematic review and meta-analysis. *BMJ: British Medical Journal*, 342, 2011.

[35] Bernard S Bloom. Crossing the quality chasm: a new health system for the 21st century. *The Journal of the American Medical Association*, 287(5):646–647, 2002.

[36] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Occam's razor. *Information processing letters*, 24(6):377–380, 1987.

[37] E Christiaan Boerma, Matty Koopmans, Arjan Konijn, Katerina Kaiferova, Andries J Bakker, Eric N van Roon, Hanneke Buter, Nienke Bruins, Peter H Egbers, Rik T Gerritsen, et al. Effects of nitroglycerin on sublingual microcirculatory blood flow in patients with severe sepsis/septic shock after a strict resuscitation protocol: a double-blind randomized placebo controlled trial. *Critical care medicine*, 38(1):93–100, 2010.

[38] RC Bone, RA Balk, FB Cerra, et al. Definition for sepsis and guide lines for the use of innovative therapies in sepsis. american colledge of chaest physician/society of critical care medicine consensus conference. *Chest*, 101:1644, 1992.

[39] Roger C Bone, Jr Charles J Fisher, Terry P Clemmer, Gus J Slotman, Craig A Metz, and Robert A Balk. Sepsis syndrome: a valid clinical entity. *Critical care medicine*, 17(5):389–393, 1989.

[40] Ailko WJ Bossink, AB Johan Groeneveld, C Erik Hack, and Lambertus G Thijs. Prediction of mortality in febrile medical patients how useful are systemic inflammatory response syndrome and sepsis criteria? *CHEST Journal*, 113(6):1533–1541, 1998.

[41] D. Peres Bota, C. Melot, F. Lopes Ferreira, V. Nguyen Ba, and J. L. Vincent. The multiple organ dysfunction score (MODS) versus the sequential organ failure assessment (SOFA) score in outcome prediction. *Intensive Care Medicine*, 28(11):1619–1624, 2002.

[42] Johan Sebastián Hernández Botero and María Cristina Florián Pérez. *The History of Sepsis from Ancient Egypt to the XIX Century*. InTech, 2012.

[43] A.P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[44] Leo Breiman. *Classification and regression trees*. CRC press, 1993.

[45] Christian Brun-Buisson, Françoise Roudot-Thoraval, Emmanuelle Girou, Catherine Grenier-Sennelier, and Isabelle Durand-Zaleski. The costs of septic syndromes in the intensive care unit and influence of hospital-acquired sepsis. *Intensive care medicine*, 29(9):1464–1471, 2003.

[46] Frank M Brunkhorst, Christoph Engel, Frank Bloos, Andreas Meier-Hellmann, Max Ragaller, Norbert Weiler, Onnen Moerer, Matthias Gruendling, Michael Oppert, Stefan Grond, et al. Intensive insulin therapy and pentastarch resuscitation in severe sepsis. *New England Journal of Medicine*, 358(2):125–139, 2008.

[47] Timothy G. Buchman, Phyllis K. Stein, and Brahm Goldstein. Heart rate variability in critical illness and critical care. *Current Opinion in Critical Care*, 8(4):311–315, Aug 2002.

[48] Bruce Campbell and Guy Maddern. Safety and efficacy of interventional procedures: scrutinising the evidence and issuing guidelines without stifling innovation. *BMJ: British Medical Journal*, 326(7385):347, 2003.

[49] Christopher R Carpenter, Samuel M Keim, Suneel Upadhye, and H Bryant Nguyen. Risk stratification of the potentially septic patient in the emergency department: the mortality in the emergency department sepsis (MEDS) score. *The Journal of emergency medicine*, 37 (3):319–327, 2009.

[50] Brendan G Carr, Jeremy M Kahn, Raina M Merchant, Andrew A Kramer, Robert W Neumar, et al. Inter-hospital variability in post-cardiac arrest mortality. *Resuscitation*, 80(1):30, 2009.

[51] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

[52] Xavier Castella, Antoni Artigas, Julian Bion, and Aarno Kari. A comparison of severity of illness scoring systems for intensive care unit patients: Results of a multicenter, multinational study. *Critical care medicine*, 23(8):1327–1335, 1995.

[53] Ferdinand Céline. *Semmelweiss*. Marbot Ediciones, 2009.

[54] Chih-Chung Chang and Chih-Jen Lin. Training v-support vector classifiers: theory and algorithms. *Neural Computation*, 13(9):2119–2147, 2001.

[55] Mary Charlson, Ted P Szatrowski, Janey Peterson, and Jeffrey Gold. Validation of a combined comorbidity index. *Journal of clinical epidemiology*, 47(11):1245–1251, 1994.

[56] Mary E Charlson, Peter Pompei, Kathy L Ales, and C Ronald MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases*, 40(5):373–383, 1987.

[57] M.E. Charlson, R.E. Charlson, J.C. Peterson, S.S. Marinopoulos, W.M. Briggs, and J.P. Hollenberg. The charlson comorbidity index is adapted to predict costs of chronic disease in primary care patients. *Journal of clinical epidemiology*, 61(12):1234–1240, 2008.

[58] Wei-Lung Chen and Cheng-Deng Kuo. Characteristics of heart rate variability can predict impending septic shock in emergency department patients with sepsis. *Academic Emergency Medicine*, 14(5):392–397, 2007.

[59] Yung-Chang Chen, Chang-Chyi Jenq, Ya-Chung Tian, Ming-Yang Chang, Chan-Yu Lin, Chih-Cheng Chang, Horng-Chyuan Lin, Ji-Tseng Fang, Chih-Wei Yang, and Shu-Min Lin. RIFLE classification for predicting in-hospital mortality in critically ill sepsis patients. *Shock*, 31(2):139–145, 2009.

[60] Joseph W Chow and Victor L Yu. Combination antibiotic therapy versus monotherapy for gram-negative bacteraemia: a commentary. *International journal of antimicrobial agents*, 11(1):7–12, 1999.

[61] Federico Cismondi, André S Fialho, Susana M Vieira, Joao MC Sousa, Shane R Reti, Leo A Celi, Michael D Howell, and Stan N Finkelstein. Predicting laboratory testing in intensive care using fuzzy and neural modeling. In *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, pages 2096–2103. IEEE, 2011.

[62] Luca Citi and Riccardo Barbieri. Physionet 2012 challenge: Predicting mortality of ICU patients using a cascaded SVM-GLM paradigm. In *Computing in Cardiology (CinC), 2012*, pages 257–260. IEEE, 2012.

[63] Gari D Clifford, William J Long, George B Moody, and Peter Szolovits. Robust parameter extraction for decision support using multimodal intensive care data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1887): 411–429, 2009.

[64] Archibald Leman Cochrane. *Effectiveness and efficiency: random reflections on health services*. Oxford: Oxford University Press, 1973, 1973.

[65] Donald W Cockcroft and M Henry Gault. Prediction of creatinine clearance from serum creatinine. *Nephron*, 16(1):31–41, 2008.

[66] *Monitoring sedation, agitation, analgesia, neuromuscular blockade, and delirium in adult ICU patients*, volume 22, 2001. Copyright© 2001 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA. Tel.:+ 1 (212) 584-4662.

[67] Ronald Cornet and Nicolette de Keizer. Forty years of snomed: a literature review. *BMC medical informatics and decision making*, 8(Suppl 1):S2, 2008.

[68] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

[69] Madalena Costa, Ary L. Goldberger, and C-K. Peng. Multiscale entropy analysis of biological signals. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 71(2 Pt 1):021906, Feb 2005.

[70] Colleen A Crowe, Erik B Kulstad, Chintan D Mistry, and Christine E Kulstad. Comparison of severity of illness scoring systems in the prediction of hospital mortality in severe sepsis and septic shock. *Journal of Emergencies, Trauma and Shock*, 3(4):342, 2010.

[71] Erik Cuevas and Miguel Cienfuegos. A new algorithm inspired in the behavior of the social-spider for constrained optimization. *Expert Systems with Applications: An International Journal*, 41(2):412–425, 2014.

[72] AA Dahaba, B Hagara, A Fall, PH Rehak, WF List, and H Metzler. Procalcitonin for early prediction of survival outcome in postoperative critically ill patients with severe sepsis. *British journal of anaesthesia*, 97(4):503–508, 2006.

[73] Margriet F C de Jong, Albertus Beishuizen, Jan-Jaap Spijkstra, and A. B Johan Groeneveld. Relative adrenal insufficiency as a predictor of disease severity, mortality, and beneficial effects of corticosteroid treatment in septic shock. *Critical Care Medicine*, 35(8):1896–1903, Aug 2007. doi: 10.1097/01.CCM.0000275387.51629.ED. URL `http://dx.doi.org/10.1097/01.CCM.0000275387.51629.ED`.

[74] R Phillip Dellinger, Jean M Carlet, Henry Masur, Herwig Gerlach, Thierry Calandra, Jonathan Cohen, Juan Gea-Banacloche, Didier Keh, John C Marshall, Margaret M Parker, et al. Surviving sepsis campaign guidelines for management of severe sepsis and septic shock. *Intensive care medicine*, 30(4):536–555, 2004.

[75] R Phillip Dellinger, Mitchell M Levy, Jean M Carlet, Julian Bion, Margaret M Parker, Roman Jaeschke, Konrad Reinhart, Derek C Angus, Christian Brun-Buisson, Richard Beale, et al. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2008. *Intensive care medicine*, 34(1):17–60, 2008.

[76] R Phillip Dellinger, Mitchell M Levy, Andrew Rhodes, Djillali Annane, Herwig Gerlach, Steven M Opal, Jonathan E Sevransky, Charles L Sprung, Ivor S Douglas, Roman Jaeschke, et al. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock, 2012. *Intensive care medicine*, 39(2):165–228, 2013.

[77] E.R. DeLong, D.M. DeLong, and D.L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

[78] J-F Dhainaut, L Thijs, and G Park. *Septic shock*. WB Saunders Company Ltd., 2000.

[79] Marco Dorigo and Mauro Birattari. Ant colony optimization. In *Encyclopedia of Machine Learning*, pages 36–39. Springer, 2010.

[80] Martin W Dünser, Jukka Takala, Hanno Ulmer, Viktoria D Mayr, Günter Luckner, Stefan Jochberger, Fritz Daudel, Philipp Lepper, Walter R Hasibeder, and Stephan M Jakob. Arterial blood pressure during early sepsis and outcome. *Intensive care medicine*, 35(7):1225–1233, 2009.

[81] Martin W Dünser, Jukka Takala, Andreas Brunauer, and Jan Bakker. Re-thinking resuscitation: leaving blood pressure cosmetics behind and moving forward to permissive hypotension and a tissue perfusion-based approach. *Critical Care*, 17:326, 2013.

[82] Moritoki Egi, Inbyung Kim, Alistair Nichol, Edward Stachowski, Craig J French, Graeme K Hart, Colin Hegarty, Michael Bailey, and Rinaldo Bellomo. Ionized calcium concentration and outcome in critical illness*. *Critical care medicine*, 39(2):314–321, 2011.

[83] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey. Comorbidity measures for use with administrative data. *Medical Care*, 36(1):8–27, Jan 1998.

[84] *Derivation and validation of automated electronic search strategies to extract charlson comorbidities from electronic medical records*, volume 87, 2012. Elsevier.

[85] E Wesley Ely, Richard Margolin, Joseph Francis, Lisa May, Brenda Truman, Robert Dittus, Theodore Speroff, Shiva Gautam, Gordon R Bernard, and Sharon K Inouye. Evaluation of delirium in critically ill patients: validation of the confusion assessment method for the intensive care unit (CAM-ICU). *Critical care medicine*, 29(7):1370–1379, 2001.

[86] E Wesley Ely, Ayumi Shintani, Brenda Truman, Theodore Speroff, Sharon M Gordon, Frank E Harrell Jr, Sharon K Inouye, Gordon R Bernard, and Robert S Dittus. Delirium as a predictor of mortality in mechanically ventilated patients in the intensive care unit. *JAMA: the journal of the American Medical Association*, 291(14):1753–1762, 2004.

[87] K. Fairchild, R. Gaykema, and L. Goehler. Heart rate variability, cytokine, and brain responses to infection: insights from a mouse model. *Critical Care*, 13:1–1, 2009.

[88] James D Faix. Biomarkers of sepsis. *Critical Reviews in Clinical Laboratory Sciences*, 50 (1):23–36, Jan 2013. doi: 10.3109/10408363.2013.764490. URL `http://dx.doi.org/10.3109/10408363.2013.764490`.

[89] Flavio Lopes Ferreira, Daliana Peres Bota, Annette Bross, Christian Mélot, and Jean-Louis Vincent. Serial evaluation of the SOFA score to predict outcome in critically ill patients. *JAMA: the journal of the American Medical Association*, 286(14):1754–1758, 2001.

[90] Mitchell P Fink and Stephen O Heard. Laboratory models of sepsis and septic shock. *Journal of Surgical Research*, 49(2):186–196, 1990.

[91] Katherine M Flegal, Brian K Kit, Heather Orpana, and Barry I Graubard. Association of all-cause mortality with overweight and obesity using standard body mass index categoriesa systematic review and meta-analysisall-cause mortality using BMI categories. *JAMA*, 309 (1):71–82, 2013.

[92] F. Fourrier, C. Chopin, J. Goudemand, S. Hendrycx, C. Caron, A. Rime, A. Marey, and P. Lestavel. Septic shock, multiple organ failure, and disseminated intravascular coagulation. compared patterns of antithrombin iii, protein c, and protein s deficiencies. *Chest*, 101(3):816, 1992.

[93] Robert A Fowler, NK Adhikari, Satish Bhagwanjee, et al. Clinical review: Critical care in the global context-disparities in burden of illness, access, and economics. *Critical Care*, 12(5): 225, 2008.

[94] I.E. Frank and J.H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, pages 109–135, 1993.

[95] Holger Frohlich, Olivier Chapelle, and Bernhard Scholkopf. Feature selection for support vector machines by means of genetic algorithm. In *Tools with Artificial Intelligence, 2003. proceedings. 15th IEEE International Conference on*, pages 142–148. IEEE, 2003.

[96] J.J. Gagne, R.J. Glynn, J. Avorn, R. Levin, and S. Schneeweiss. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *Journal of Clinical Epidemiology*, 64:pp749–759, 0802Jan 2011. doi: 10.1016/j.jclinepi.2010.10.004. URL `http://dx.doi.org/10.1016/j.jclinepi.2010.10.004`.

[97] J.r.l.e. Gall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, and D. Villers. A simplified acute physiology score for ICU patients. *Critical Care Medicine*, 12(11):975, 1984. ISSN 0090-3493.

[98] Yi Gang and Marek Malik. Heart rate variability in critical care medicine. *Current Opinion in Critical Care*, 8(5):371–375, Oct 2002.

[99] Luciano Gattinoni, Luca Brazzi, Paolo Pelosi, Roberto Latini, Gianni Tognoni, Antonio Pesenti, and Roberto Fumagalli. A trial of goal-oriented hemodynamic therapy in critically ill patients. *New England Journal of Medicine*, 333(16):1025–1032, 1995.

[100] Atul Gawande. The checklist. *The New Yorker*, 83(39):86–95, 2007.

[101] H F Geerdes, D Ziegler, H Lode, M Hund, A Loehr, W Fangmann, and J Wagner. Septicemia in 980 patients at a university hospital in berlin: prospective studies during 4 selected years between 1979 and 1989. *Clinical Infectious Diseases*, 15(6):991–1002, Dec 1992.

[102] Ming Geng. *A comparison of logistic regression to random forests for exploring differences in risk factors associated with stage atdiagnosis between black and white colon cancer patients*. PhD thesis, University of Pittsburgh, 2006.

[103] Stefanos Geroulanos and Evangelia T Douka. Historical perspective of the word "sepsis". *Intensive care medicine*, 32(12):2077–2077, 2006.

[104] Marzyeh Ghassemi. Methods and models for acute hypotensive episode prediction. Master's thesis, University of Oxford, Linacre College, 2010.

[105] Timothy D Girard, Pratik P Pandharipande, and E Wesley Ely. Delirium in the intensive care unit. *Critical Care*, 12(Suppl 3):S3, 2008.

[106] DE Goldberg et al. *Genetic Algorithms In Search, Optimization, And Machine Learning*. Addison-wesley Reading Menlo Park, 1989.

[107] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, and H.E. Stanley. Physiobank, Pysiotoolkit, and Physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23): e215, 2000.

[108] Ronald D Gonzales, Paul C Schreckenberger, Mary Beth Graham, Swathi Kelkar, Karen DenBesten, and John P Quinn. Infections due to vancomycin-resistant< i> enterococcus faecium</i> resistant to linezolid. *The Lancet*, 357(9263):1179, 2001.

[109] Neville W Goodman. Ethics and evidence-based medicine: Fallibility and responsibility in clinical science. *JRSM*, 96(5):251–251, 2003.

[110] Bertrand Guidet, Olivier Martinet, Thierry Boulain, Francois Philippart, Jean F Poussel, Julien Maizel, Xavier Forceville, Marc Feissel, Michel Hasselmann, Alexandra Heininger, et al. Assessment of hemodynamic efficacy and safety of 6% hydroxyethylstarch 130/0.4 vs. 0.9% nacl fluid replacement in patients with severe sepsis: The crystmas study. *Critical Care*, 16 (3):R94, 2012.

[111] Jose Gutierrez, Hossam Amin, Roxana Lazareacu, El Kay, and Tatjana Rundek. Effect of beta blockers on sepsis outcome. *Medical science monitor*, 15(10), 2009.

[112] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[113] J J Haddad and HL Harb. Cytokines and the regulation of hypoxia-inducible factor (hif)-1alpha. *International Immunopharmacology*, 5(3):461–483, Mar 2005. doi: 10.1016/j.intimp.2004.11.009. URL `http://dx.doi.org/10.1016/j.intimp.2004.11.009`.

[114] Edward O Hahn, Harold B Houser, Charles H Rammelkamp Jr, Floyd W Denny, and Lewis W Wannamaker. Effect of cortisone on acute streptococcal infections and post-streptococcal complications. *Journal of Clinical Investigation*, 30(3):274, 1951.

[115] Neil A Halpern and Stephen M Pastores. Critical care medicine in the United States 2000-2005: An analysis of bed numbers, occupancy rates, payer mix, and costs*. *Critical care medicine*, 38(1):65–71, 2010.

[116] Neil A Halpern, Stephen M Pastores, and Robert J Greenstein. Critical care medicine in the United States 1985–2000: An analysis of bed numbers, use, and costs*. *Critical care medicine*, 32(6):1254–1259, 2004.

[117] Pat Hamilton. Open source ecg analysis. In *Computers in Cardiology, 2002*, pages 101–104. IEEE, 2002.

[118] JA Hanley and BJ McNeil. The meaning and use of the area under a receiver operating (ROC) curvel characteristic. *Radiology*, 143(1):29–36, 1982.

[119] J.A. Hanley, B.J. McNeil, et al. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843, 1983.

[120] Frank E Harrell. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, 2001.

[121] Per-Olof Hasselgren, J Howard James, Daniel W Benson, Marianne Hall-Angerås, Ulf Angerås, Darryl T Hiyama, Shujun Li, and Josef E Fischer. Total and myofibrillar protein breakdown in different types of rat skeletal muscle: effects of sepsis and regulation by insulin. *Metabolism: Clinical and Experimental*, 38(7):634–640, 1989.

[122] Reinhold Haux. Health information systems? past, present, future. *International Journal of Medical Informatics*, 75(3-4):268–281, 2006.

[123] Reinhold Haux, Alfred Winter, Elske Ammenwerth, and Birgit Brigl. *Strategic information management in hospitals: an introduction to hospital information systems*. Springer, 2004.

[124] George B Haycock, George J Schwartz, and David H Wisotsky. Geometric method for measuring body surface area: a height-weight formula validated in infants, children, and adults. *The Journal of pediatrics*, 93(1):62–66, 1978.

[125] Michelle A Hayes, Andrew C Timmins, Ernest Yau, Mark Palazzo, Charles J Hinds, and David Watson. Elevation of systemic oxygen delivery in the treatment of critically ill patients. *New England Journal of Medicine*, 330(24):1717–1722, 1994.

[126] Alex B Haynes, Thomas G Weiser, William R Berry, Stuart R Lipsitz, Abdel-Hadi S Breizat, E Patchen Dellinger, Teodoro Herbosa, Sudhir Joseph, Pascience L Kibatala, Marie Carmela M Lapitan, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. *New England Journal of Medicine*, 360(5):491–499, 2009.

[127] Paul C Hebert, AJ Drummond, J Singer, GR Bernard, and JA Russell. A simple multiple system organ failure scoring system predicts mortality of patients who have sepsis syndrome. *CHEST Journal*, 104(1):230–235, 1993.

[128] Paul C Hébert, George Wells, Morris A Blajchman, John Marshall, Claudio Martin, Giuseppe Pagliarello, Martin Tweeddale, Irwin Schweitzer, and Elizabeth Yetisir. A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. *New England Journal of Medicine*, 340(6):409–417, 1999.

[129] Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. Randomized trials analyzed as observational studies. *Annals of internal medicine*, 159(8):560–562, 2013.

[130] Thomas L Higgins, Daniel Teres, Wayne S Copes, Brian H Nathanson, Maureen Stark, and Andrew A Kramer. Assessing contemporary intensive care unit outcome: An updated mortality probability admission model (mpm0-iii)*. *Critical care medicine*, 35(3):827–835, 2007.

[131] T.L. Higgins, D. Teres, W. Copes, B. Nathanson, M. Stark, and A. Kramer. Updated mortality probability model (MPM-iii). *CHEST Journal*, 128(4_MeetingAbstracts):348S–348S, 2005.

[132] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, pages 55–67, 1970.

[133] Cheryl L Holmes, James A Russell, and Keith R Walley. Genetic polymorphisms in sepsis and septic shock: role in prognosis and potential for therapy. *Chest*, 124(3):1103–1115, Sep 2003.

[134] Ewout J Hoorn, Michiel GH Betjes, Joachim Weigel, and Robert Zietse. Hypernatraemia in critically ill patients: too little water and too much salt. *Nephrology Dialysis Transplantation*, 23(5):1562–1568, 2008.

[135] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. Wiley. com, 2013.

[136] R. S. Hotchkiss, P. E. Swanson, B. D. Freeman, K. W. Tinsley, J. P. Cobb, G. M. Matuschak, T. G. Buchman, and I. E. Karl. Apoptotic cell death in patients with sepsis, shock, and multiple organ dysfunction. *Critical Care Medicine*, 27(7):1230–1251, Jul 1999.

[137] C.R. Houck, J. Joines, and M. Kay. A genetic algorithm for function optimization: A matlab implementation. *NCSU-IE TR*, 95(09), 1995.

[138] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification, 2003.

[139] Bo Hu, Srinandan Dasmahapatra, David Dupplaw, Paul Lewis, and Nigel Shadbolt. Reflections on a medical ontology. *International journal of human-computer studies*, 65(7): 569–582, 2007.

[140] Cheng-Lung Huang and Chieh-Jen Wang. A ga-based feature selection and parameters optimizationfor support vector machines. *Expert Systems with applications*, 31(2):231–240, 2006.

[141] Caleb W Hug, Gari D Clifford, and Andrew T Reisner. Clinician blood pressure documentation of stable intensive care patients: an intelligent archiving agent has a higher association with future hypotension. *Critical Care Medicine*, 39(5):1006–1014, May 2011. doi: 10.1097/CCM.0b013e31820eab8e. URL `http://dx.doi.org/10.1097/CCM.0b013e31820eab8e`.

[142] Tsann-Long Hwang. Sex different responses and immunomodulation in severe sepsis. *Formosan Journal of Surgery*, 2012.

[143] *Predicting in-hospital mortality of ICU patients: The Physionet/Computing in cardiology challenge 2012*, Computing in Cardiology (CinC), 2012, 2012. IEEE.

[144] M.F. Jefferson, N. Pendleton, S.B. Lucas, and M.A. Horan. Comparison of a genetic algorithm neural network with logistic regression for predicting outcome after surgery for patients with nonsmall cell lung carcinoma. *Cancer*, 79(7):1338–1342, 1997.

[145] Alistair EW Johnson, Nic Dunkley, Louis Mayaud, Athanasios Tsanas, Andrew A Kramer, and Gari D Clifford. Patient specific predictions in the intensive care unit using a bayesian ensemble. *Computer in Cardiology*, pages 249–252, 2012.

[146] Alistair EW Johnson, Andrew A Kramer, and Gari D Clifford. A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy*. *Critical care medicine*, 41(7):1711–1718, 2013.

[147] Alan E Jones, Stephen Trzeciak, and Jeffrey A Kline. The sequential organ failure assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation. *Critical care medicine*, 37(5):1649, 2009.

[148] Leo Anthony Celi Joon Lee. The obesity paradox in the ICU is also likely an artifact. Personal communication, 2013.

[149] Dervis Karaboga. An idea based on honey bee swarm for numerical optimization. *Techn. Rep. TR06, Erciyes Univ. Press, Erciyes*, 2005.

[150] Stylianos Katsaragakis, Konstantinos Papadimitropoulos, Pantelis Antonakis, Spyros Strergiopoulos, Manoussos M Konstadoulakis, and George Androulakis. Comparison of acute physiology and chronic health evaluation II (APACHE II) and simplified acute physiology score II (SAPS II) scoring systems in a single greek intensive care unit. *Critical care medicine*, 28(2):426–432, 2000.

[151] Taro Kawai and Shizuo Akira. Innate immune recognition of viral infection. *Nature Immunology*, 7(2):131–137, Feb 2006. doi: 10.1038/ni1303. URL `http://dx.doi.org/10.1038/ni1303`.

[152] Mark T Keegan, Ognjen Gajic, and Bekele Afessa. Comparison of APACHE iii and iv, SAPS 3 and mpm0iii, and influence of resuscitation status on model performance. *CHEST Journal*, 2012.

[153] Mark A Kelley, Derek Angus, Donald B Chalfin, Edward D Crandall, David Ingbar, Wanda Johanson, Justine Medina, Curtis N Sessler, and Jeffery S Vender. The critical care crisis in the united statesa report from the profession. *CHEST Journal*, 125(4):1514–1517, 2004.

[154] James Kennedy. Particle swarm optimization. In *Encyclopedia of Machine Learning*, pages 760–766. Springer, 2010.

[155] Michelle M Kim, Amber E Barnato, Derek C Angus, Lee F Fleisher, and Jeremy M Kahn. The effect of multidisciplinary care teams on intensive care unit mortality. *Archives of internal medicine*, 170(4):369, 2010.

[156] Kenji Kira and Larry A Rendell. A practical approach to feature selection. In *proceedings of the ninth international workshop on Machine learning*, pages 249–256. Morgan Kaufmann Publishers Inc., 1992.

[157] Zosia Kmietowicz. Half of patients in intensive care receive suboptimal care. *BMJ: British Medical Journal*, 330(7500):1101, 2005.

[158] W.A. Knaus, DP Wagner, EA Draper, JE Zimmerman, M. Bergner, PG Bastos, CA Sirio, DJ Murphy, T. Lotring, and A. Damiano. The APACHE iii prognostic system. risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*, 100(6):1619, 1991.

[159] William A Knaus, Jack E Zimmerman, Douglas P Wagner, Elizabeth A Draper, and Diane E Lawrence. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Critical care medicine*, 9(8):591–597, 1981.

[160] William A Knaus, Elizabeth A Draper, Douglas P Wagner, and Jack E Zimmerman. APACHE II: a severity of disease classification system. *Critical care medicine*, 13(10):818–829, 1985.

[161] William A Knaus, Douglas P Wagner, and Joanne Lynn. Short-term mortality predictions for critically ill hospitalized adults: science and ethics. *Science*, 254(5030):389–394, 1991.

[162] William A Knaus, Frank E Harrell, Charles J Fisher Jr, Douglas P Wagner, Steven M Opal, Jerald C Sadoff, Elizabeth A Draper, Cynthia A Walawander, Kathleen Conboy, and Thaddeus H Grasela. The clinical evaluation of new drugs for sepsis. *JAMA: the journal of the American Medical Association*, 270(10):1233–1241, 1993.

[163] William A. Knaus, Douglas P. Wagner, Jack E. Zimmerman, and Elizabeth A. Draper. Variations in mortality and length of stay in intensive care units. *Annals of Internal Medicine*, 118(10):753–761, 1993. doi: 10.7326/0003-4819-118-10-199305150-00001. URL `+http://dx.doi.org/10.7326/0003-4819-118-10-199305150-00001`.

[164] Igor Kononenko, Edvard Šimec, and Marko Robnik-Šikonja. Overcoming the myopia of inductive learning algorithms with relieff. *Applied Intelligence*, 7(1):39–55, 1997.

[165] Dimitrios A Kontoyannis and Massimo Piepoli. Spectral analysis of heart rate variability in the sepsis syndrome. *Clinical Autonomic Research*, 3(1):5–13, 1993.

[166] Rishi Kothari, Joseph Ladapo, Daniel Scott, and Leo Celi. Interrogating a clinical database to study treatment of hypotension in the critically ill. *BMJ Open*, 2:10.1136/bmjopen-2012-000916, 2012.

[167] Andrew A Kramer and Jack E Zimmerman. Assessing the calibration of mortality benchmarks in critical care: The hosmer-lemeshow test revisited*. *Critical care medicine*, 35(9): 2052–2056, 2007.

[168] Wojtek J Krzanowski and WJ Krzanowski. *Principles of multivariate analysis*. Oxford University Press, 1996.

[169] I Kwan, F Bunn, I Roberts, WHO Pre-Hospital Trauma Care Steering Committee, et al. Timing and volume of fluid administration for patients with bleeding. *Cochrane Database Syst Rev*, 3(3), 2003.

[170] Douglas E Lake, Joshua S Richman, M Pamela Griffin, and J Randall Moorman. Sample entropy analysis of neonatal heart rate variability. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 283(3):R789–R797, 2002.

[171] Wesley H Self Laurie A. Hawkins. Multicenter blood culture quality improvement. Technical report, 2011. URL `http://www.clinicaltrials.gov/ct2/show/NCT01413555`.

[172] *A study of cross-validation and bootstrap for accuracy estimation and model selection*, volume 14, 1995. Lawrence Erlbaum Associates Ltd.

[173] Jean-Roger Le Gall, Philippe Loirat, Annick Alperovitch, Paul Glaser, Claude Granthil, Daniel Mathieu, Philippe Mercier, Remi Thomas, and Daniel Villers. A simplified acute physiology score for ICU patients. *Critical care medicine*, 12(11):975–977, 1984.

[174] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (SAPS II) based on a european/north american multicenter study. *JAMA: the journal of the American Medical Association*, 270(24):2957–2963, 1993.

[175] J Lee and RG Mark. An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care. Rev4, 2010.

[176] S. Lemeshow and D.W. Hosmer. A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology*, 115(1):92, 1982.

[177] Stanley Lemeshow, Daniel Teres, Harris Pastides, Jill Spitz Avrunin, and Jay S Steingrub. A method for predicting survival and mortality of ICU patients using objectively derived weights. *Critical care medicine*, 13(7):519–525, 1985.

[178] StanleyLEY Lemeshow, Daniel Teres, Jill Spitz Avrunin, and Robert W Gage. Refining intensive care unit outcome prediction by using changing probabilities of mortality. *Critical care medicine*, 16(5):470–477, 1988.

[179] Mitchell M Levy, R Phillip Dellinger, Sean R Townsend, Walter T Linde-Zwirble, John C Marshall, Julian Bion, Christa Schorr, Antonio Artigas, Graham Ramsay, Richard Beale, et al. The surviving sepsis campaign: results of an international guideline-based performance improvement program targeting severe sepsis. *Intensive care medicine*, 36(2):222–231, 2010.

[180] M.M. Levy, M.P. Fink, J.C. Marshall, E. Abraham, D. Angus, D. Cook, J. Cohen, S.M. Opal, J.L. Vincent, and G. Ramsay. Sepsis definition. In *2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference*, volume 29, pages 530–538. Springer, 2003.

[181] Li Li, Wei Jiang, Xia Li, Kathy L Moser, Zheng Guo, Lei Du, Qiuju Wang, Eric J Topol, Qing Wang, and Shaoqi Rao. A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. *Genomics*, 85(1):16–23, 2005.

[182] Hsuan-Tien Lin and Chih-Jen Lin. A study on sigmoid kernels for SVM and the training of non-psd kernels by smo-type methods. *submitted to Neural Computation*, pages 1–32, 2003.

[183] Shih-Wei Lin, Zne-Jung Lee, Shih-Chieh Chen, and Tsung-Yuan Tseng. Parameter determination of support vector machine and feature selection using simulated annealing approach. *Applied soft computing*, 8(4):1505–1512, 2008.

[184] L. M. Lix, J. Quail, G. Teare, and B. Acan. Performance of comorbidity measures for predicting outcomes in population-based osteoporosis cohorts. *Osteoporosis International*, Jan 2011. doi: 10.1007/s00198-010-1516-7. URL `http://dx.doi.org/10.1007/s00198-010-1516-7`.

[185] Kox M and Pickkers P. "less is more" in critically ill patients: Not too intensive. *JAMA Internal Medicine*, 173(14):1369–1372, 2013. doi: 10.1001/jamainternmed.2013.6702. URL `+http://dx.doi.org/10.1001/jamainternmed.2013.6702`.

[186] D.J.C. MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[187] D.J.C. MacKay. Probable networks and plausible predictions-a review of practical bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3): 469–505, 1995.

[188] D.J.C. MacKay et al. Bayesian nonlinear modeling for the prediction competition. *ASHRAE transactions*, 100(2):1053–1062, 1994.

[189] Greg S. Martin, David M. Mannino, Stephanie Eaton, and Marc Moss. The epidemiology of sepsis in the United States from 1979 through 2000. *New England Journal of Medicine*, 348(16):1546–1554, Apr 2003. doi: 10.1056/NEJMoa022139. URL `http://dx.doi.org/10.1056/NEJMoa022139`.

[190] I. Matot and C.L. Sprung. Definition of sepsis. *Intensive Care Medicine*, 27(14):3–9, 2001.

[191] Louis Mayaud, Peggy S Lai, Gari D Clifford, Lionel Tarassenko, Leo Anthony Celi, and Djillali Annane. Dynamic data during hypotensive episode improves mortality predictions among patients with sepsis and hypotension*. *Critical care medicine*, 41(4):954–962, 2013.

[192] Louis Mayaud, Lionel Tarassenko, Djillali Annane, and Gari Clifford. Predictive power of heart rate complexity to estimate severity in severe sepsis patients. *Journal of Critical Care*, 28:e37, 2013.

[193] Clement J McDonald, Stanley M Huff, Jeffrey G Suico, Gilbert Hill, Dennis Leavelle, Raymond Aller, Arden Forrey, Kathy Mercer, Georges DeMoor, John Hook, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clinical chemistry*, 49 (4):624–633, 2003.

[194] JE Jr McGowan, PL Parrott, and VP Duty. Nosocomial bacteriamea. potential for prevention of procedure related case. *JAMA*, 237:2727–2729, 1977.

[195] Kresten Mellemgaard. The alveolar-arterial oxygen difference: Its size and components in normal man. *Acta Physiologica Scandinavica*, 67(1):10–20, 1966.

[196] Marvin Lee Minsky. *Heuristic aspects of the artificial intelligence problem*. Massachusetts Institute of Technology, Lincoln Laboratory, 1956.

[197] George B Moody and Roger G Mark. Development and evaluation of a 2-lead ecg analysis program. *Computers in Cardiology*, 9:39–44, 1982.

[198] George B Moody and Roger G Mark. The MIT-BIH arrhythmia database on cd-rom and software for use with it. In *Computers in Cardiology 1990, proceedings.*, pages 185–188. IEEE, 1990.

[199] George B Moody, Roger G Mark, and Ary L Goldberger. Physionet: a web-based resource for the study of physiologic signals. *Engineering in Medicine and Biology Magazine, IEEE*, 20(3):70–75, 2001.

[200] Leonardo A Moraes, Mark J Paul-Clark, Alice Rickman, Roderick J Flower, Nicolas J Goulding, and Mauro Perretti. Ligand-specific glucocorticoid receptor activation in human platelets. *Blood*, 106(13):4167–4175, 2005.

[201] R Moreno, B Jordan, and P Metnitz. The changing prognostic determinants in the critically iii patient. In *Intensive Care Medicine*, pages 899–907. Springer, 2007.

[202] Rui Moreno and Giovanni Apolone. Impact of different customization strategies in the performance of a general severity score. *Critical care medicine*, 25(12):2001–2008, 1997.

[203] Rui P Moreno, Philipp GH Metnitz, Eduardo Almeida, Barbara Jordan, Peter Bauer, Ricardo Abizanda Campos, Gaetano Iapichino, David Edbrooke, Maurizia Capuzzo, and Jean-Roger Le Gall. SAPS 3—from evaluation of the patient to evaluation of the intensive care unit. part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive care medicine*, 31(10):1345–1355, 2005.

[204] Rui P Moreno, Barbara Metnitz, Leopold Adler, Anette Hoechtl, Peter Bauer, and Philipp GH Metnitz. Sepsis mortality prediction based on predisposition, infection and response. *Intensive care medicine*, 34(3):496–504, 2008.

[205] H. Muhlenbein. Evolution in time and space-the parallel genetic algorithm. In *Foundations of genetic algorithms*. Citeseer, 1991.

[206] Klaus-Robert Muller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, and Bernhard Scholkopf. An introduction to kernel-based learning algorithms. *IEEE transactions on neural networks*, 12(2):181–201, 2001.

[207] Maryam Nejat, John W Pickering, Robert J Walker, Justin Westhuyzen, Geoffrey M Shaw, Christopher M Frampton, and Zoltán H Endre. Urinary cystatin c is diagnostic of acute kidney injury and sepsis, and predicts mortality in the intensive care unit. *Critical Care*, 14 (3):1–13, 2010.

[208] Shamim Nemati, Li-wei H. Lehman, Louis Mayaud, G. Clifford, and R. P. Adams. Time series dynamics and patient state monitoring: Application to ICU bedside predictive analytics. In *Machine Learning for Clinical Data Analysis and Healthcare, NIPS Workshop 2013*, 2013.

[209] C.M. Norris, W.A. Ghali, M.L. Knudtson, C.D. Naylor, and L.D. Saunders. Dealing with missing data in observational health care outcome analyses. *Journal of clinical epidemiology*, 53(4): 377–383, 2000.

[210] MAFP Novaes, A Aronovich, MB Ferraz, and E Knobel. Stressors in ICU: patients' evaluation. *Intensive Care Medicine*, 23(12):1282–1285, 1997.

[211] Nathan M Novotny, Tim Lahm, Troy A Markel, Paul R Crisostomo, Meijing Wang, Yue Wang, Rinki Ray, Jiangning Tan, Dalia Al-Azzawi, and Daniel R Meldrum. Beta-blockers in sepsis: reexamining the evidence. *Shock*, 31(2):113–119, 2009.

[212] Kimberly J O'Malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. Measuring diagnoses: ICD code accuracy. *Health services research*, 40(5p2):1620–1639, 2005.

[213] M Oppert, C Engel, FM Brunkhorst, H Bogatsch, K Reinhart, U Frei, KU Eckardt, M Loeffler, and S John. German competence network sepsis (Sepnet). acute renal failure in patients with severe sepsis and septic shock, a significant independent risk factor for mortality: results from the german prevalence study. *Nephrology, Dialysis, Transplantation*, 23(3): 904–9, 2008.

[214] Gustavo A Ospina-Tascón, Gustavo Luiz Büchele, and Jean-Louis Vincent. Multicenter, randomized, controlled trials evaluating mortality in intensive care: Doomed to fail? *Critical care medicine*, 36(4):1311–1322, 2008.

[215] Andrew Padkin, Caroline Goldfrad, Anthony R Brady, Duncan Young, Nick Black, and Kathy Rowan. Epidemiology of severe sepsis occurring in the first 24 hrs in intensive care units in england, wales, and northern ireland. *Critical care medicine*, 31(9):2332–2338, 2003.

[216] Ping-Feng Pai and Wei-Chiang Hong. Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms. *Electric Power Systems Research*, 74 (3):417–425, 2005.

[217] Jiapu Pan and Willis J Tompkins. A real-time qrs detection algorithm. *Biomedical Engineering, IEEE Transactions on*, 32(3):230–236, 1985.

[218] J.E. Parrillo, M.M. Parker, C. Natanson, A.F. Suffredini, R.L. Danner, R.E. Cunnion, and F.P. Ognibene. Septic shock in humans. *Annals of Internal Medicine*, 113(3):227, 1990.

[219] Stephen M Pastores, Jubran Dakwar, and Neil A Halpern. Costs of critical care medicine. *Critical Care Clinics*, 28(1):1–10, 2012.

[220] Gourang P Patel and Robert A Balk. Systemic steroids in severe sepsis and septic shock. *American journal of respiratory and critical care medicine*, 185(2):133–139, 2012.

[221] M.J. Pencina, R.B. D'Agostino Sr, R.B. D'Agostino Jr, and R.S. Vasan. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine*, 27(2):157–172, 2008.

[222] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell*, pages 1226–1238, 2005. ISSN 0162-8828.

[223] Anders Perner, Nicolai Haase, Anne B Guttormsen, Jyrki Tenhunen, Gudmundur Klemenzson, Anders Åneman, Kristian R Madsen, Morten H Møller, Jeanie M Elkjær, Lone M Poulsen, et al. Hydroxyethyl starch 130/0.42 versus ringer's acetate in severe sepsis. *New England Journal of Medicine*, 367(2):124–134, 2012.

[224] Charalampos Pierrakos and Jean-Louis Vincent. Sepsis biomarkers: a review. *Critical Care*, 14(1):R15, 2010. doi: 10.1186/cc8872. URL `http://dx.doi.org/10.1186/cc8872`.

[225] Didier Pittet, Bernard Thiévent, Richard P Wenzel, Ning Li, Raymond Auckenthaler, and Peter M Suter. Bedside prediction of mortality from bacteremic sepsis. a dynamic analysis of ICU patients. *American journal of respiratory and critical care medicine*, 153(2):684–693, 1996.

[226] K. Pollak and A.N. Sauquet. *Los discípulos de Hipócrates: una historia de la medicina*. Enciclopedia general de la cultura. Círculo de Lectores, 1969. URL `http://books.google.co.uk/books?id=KWSnmQEACAAJ`.

[227] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C: The Art of scientific Computing*. Cambridge University Press, 1992.

[228] Peter Pronovost, Dale Needham, Sean Berenholtz, David Sinopoli, Haitao Chu, Sara Cosgrove, Bryan Sexton, Robert Hyzy, Robert Welsh, Gary Roth, et al. An intervention to decrease catheter-related bloodstream infections in the ICU. *New England Journal of Medicine*, 355(26):2725–2732, 2006.

[229] Hude Quan, Gerry A Parsons, and William A Ghali. Validity of information on comorbidity derived from ICD-9-ccm administrative data. *Medical care*, 40(8):675–685, 2002.

[230] A. A. Quartin, R. M. Schein, D. H. Kett, and P. N. Peduzzi. Magnitude and duration of the effect of sepsis on survival. department of veterans affairs systemic sepsis cooperative studies group. *JAMA*, 277(13):1058–1063, Apr 1997.

[231] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[232] J. Ross Quinlan. Improved use of continuous attributes in c4. 5. *arXiv preprint cs/9603103*, 1996.

[233] Mohamed Y Rady. The role of central venous oximetry, lactic acid concentration and shock index in the evaluation of clinical shock: a review. *Resuscitation*, 24(1):55–60, 1992.

[234] Amir H Sadrzadeh Rafie, Frederick E Dewey, Gannon W Sungar, Euan A Ashley, David Hadley, Jonathan Myers, and Victor F Froelicher. Age and double product (systolic blood pressure$\times$ heart rate) reserve-adjusted modification of the duke treadmill score nomogram in men. *The American Journal of Cardiology*, 102(10):1407–1412, 2008.

[235] M Rapin, G Gory, F Lemaire, B Teisseire, and A Harari. Haemodynamic effects of dopamine in septic shock. *Intensive care medicine*, 3(2):47–53, 1977.

[236] Jordi Rello, Alejandro Rodriguez, Thiago Lisboa, Miguel Gallego, Manel Lujan, and Richard Wunderink. PIRO score for community-acquired pneumonia: A new prediction rule for assessment of severity in intensive care unit patients with community-acquired pneumonia*. *Critical care medicine*, 37(2):456–462, 2009.

[237] Ji-Young Rhee, Ki Tae Kwon, Hyun Kyun Ki, Sang Yop Shin, Dong Sik Jung, Doo-Ryeon Chung, Byoung-Chun Ha, Kyong Ran Peck, and Jae-Hoon Song. Scoring systems for prediction of mortality in patients with intensive care unit-acquired sepsis: a comparison of the pitt bacteremia score and the acute physiology and chronic health evaluation II scoring systems. *Shock*, 31(2):146–150, 2009.

[238] Andrew Rhodes and E David Bennett. Early goal-directed therapy: an evidence-based review. *Critical care medicine*, 32(11):S448–S450, 2004.

[239] Andrew Rhodes, Fiona J Lamb, Ignazio Malagon, Philip J Newman, R Michael Grounds, and E David Bennett. A prospective study of the use of a dobutamine stress test to identify outcome in patients with sepsis, severe sepsis, or septic shock. *Critical care medicine*, 27 (11):2361–2366, 1999.

[240] Lu Richard and Ronilda Lacson. Mortality prediction in patients with septic shock: an updated logistic model and a comparison with other classification algorithms. Unpublished, dicklu@gmail.com, 2008.

[241] Richard R Riker, Jean T Picard, and Gilles L Fraser. Prospective evaluation of the sedation-agitation scale for adult critically ill patients. *Critical care medicine*, 27(7):1325–1329, 1999.

[242] William P Riordan, Patrick R Norris, Judith M Jenkins, and John A Morris. Early loss of heart rate complexity predicts mortality regardless of mechanism, anatomic location, or severity of injury in 2178 trauma patients. *The Journal of surgical research*, 156(2):283–289, 2009.

[243] E. Rivers, B. Nguyen, S. Havstad, J. Ressler, A. Muzzin, B. Knoblich, E. Peterson, M. Tomlanovich, et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *New England Journal of Medicine*, 345(19):1368, 2001.

[244] Emanuel P Rivers, H Bryant Nguyen, David T Huang, and Michael Donnino. Early goal-directed therapy. *Critical care medicine*, 32(1):314–315, 2004.

[245] JD Robinson, SM Lupkiewicz, L Palenik, LM Lopez, and M Ariet. Determination of ideal body weight for drug dosage calculations. *American Journal of Health-System Pharmacy*, 40(6):1016–1019, 1983.

[246] Teemu Roos, Hannes Wettig, Peter Grünwald, Petri Myllymäki, and Henry Tirri. On discriminative bayesian network classifiers and logistic regression. *Machine Learning*, 59(3): 267–296, 2005.

[247] James A Russell. Management of sepsis. *New England Journal of Medicine*, 355(16): 1699–1713, 2006.

[248] Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. *Critical Care Medicine*, Jan 2011. doi: 10.1097/CCM.0b013e31820a92c6. URL `http://dx.doi.org/10.1097/CCM.0b013e31820a92c6`.

[249] Y Sakr, C Krauss, ACKB Amaral, A Rea-Neto, M Specht, K Reinhart, and G Marx. Comparison of the performance of SAPS II, SAPS 3, APACHE II, and their customized prognostic models in a surgical intensive care unit. *British journal of anaesthesia*, 101(6):798–803, 2008.

[250] Elmar Schlich, M Schumm, and M Schlich. 3-d-body-scan als anthropometrisches verfahren zur bestimmung der spezifischen körperoberfläche. *Ernährungs Umschau*, 57:178–183, 2010.

[251] Bernhard Schölkopf, Alex J Smola, Robert C Williamson, and Peter L Bartlett. New support vector algorithms. *Neural computation*, 12(5):1207–1245, 2000.

[252] RB Shafer and JA Bianco. Assessment of cardiac reserve in patients with hyperthyroidism. *CHEST Journal*, 78(2):269–273, 1980.

[253] E. Shannon and W Weaver. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[254] Nathan I Shapiro, Richard E Wolfe, Richard B Moore, Eric Smith, Elizabeth Burdick, and David W Bates. Mortality in emergency department sepsis (MEDS) score: A prospectively derived and validated clinical prediction rule*. *Critical care medicine*, 31(3):670–675, 2003.

[255] D. Shavdia. Septic shock: providing early warnings through multivariate logistic regression models. Master's thesis, Massachusetts Institute of Technology, 2007.

[256] Herbert Shubin and Max H Weil. Bacterial shock. *JAMA: The Journal of the American Medical Association*, 185(11):850–853, 1963.

[257] Herbert Shubin and Max Harry Weil. Bacterial shock. *JAMA: the journal of the American Medical Association*, 235(4):421–424, 1976.

[258] C. A. Siro, P. G. Bastos, W. A. Knaus, and D. P. Wagner. APACHE II scores in the prediction of multiple organ failure syndrome. *Archives of Surgery*, 126(4):528–529, Apr 1991.

[259] Elizabeth Slade, Pritpal S Tamber, and Jean-Louis Vincent. The surviving sepsis campaign: raising awareness to reduce mortality. *Critical Care*, 7(1):1, 2003.

[260] Erick D Slazinski. Structured query language (sql). *The Internet Encyclopedia*, 2004.

[261] Crowd source. Half of patients in intensive care receive suboptimal care - recent rapid responses. http://www.BMJ.com/content/330/7500/1101.1?tab=responses, 2005.

[262] David H Spodick. The swan-ganz catheterrequesting scientific trials is not an "assault". *CHEST Journal*, 115(3):857–858, 1999.

[263] Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ: British Medical Journal*, 338, 2009.

[264] Bowman Sue. Coordinating snomed-ct and ICD-10: Getting the most out of electronic health record systems. *Perspectives in Health Information Management White paper*, 20050526, 2005.

[265] J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[266] JAK Suykens, T Van Gestel, J De Brabanter, B De Moor, and J Vandewalle. Support vector machine. *World Scientific*, 2002.

[267] Task-Force. Heart-rate variability: Standard of measurment, physiological interpretation and clinical use. *European Heart Journal*, 17:354–381, 1996.

[268] Graham Teasdale and Bryan Jennett. Assessment of coma and impaired consciousness: a practical scale. *The Lancet*, 304(7872):81–84, 1974.

[269] Daniel Teres, Stanley Lemeshow, JILL SPITZ AVRUNIN, and HARRIS PASTIDES. Validation of the mortality prediction model for ICU patients. *Critical care medicine*, 15(3):208–213, 1987.

[270] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 0035-9246.

[271] R. Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16:385–396, 1997. URL `http://www-stat.stanford.edu/~tibs/lasso/fulltext.pdf`.

[272] M.E. Tipping. Sparse bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research*, 1:211–244, 2001.

[273] Cheng Toh, Lawrence O Ticknor, Colin Downey, Alan R Giles, Ray C Paton, and Richard Wenstone. Early identification of sepsis and mortality risks through simple, rapid clot-waveform analysis. *Intensive care medicine*, 29(1):55–61, 2003.

[274] Athanasios Tsanas, Max A Little, Patrick E McSharry, Jennifer Spielman, and Lorraine O Ramig. Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease. *Biomedical Engineering, IEEE Transactions on*, 59(5):1264–1271, 2012.

[275] Sotirios Tsiodras, Howard S Gold, George Sakoulas, George M Eliopoulos, Christine Wennersten, Lata Venkataraman, Robert C Moellering Jr, and Mary Jane Ferraro. Linezolid resistance in a clinical isolate of staphylococcus aureus. *The Lancet*, 358(9277):207–208, 2001.

[276] KJ Tuman, RJ McCarthy, RJ March, H Najafi, and AD Ivankovich. Morbidity and duration of ICU stay after cardiac surgery. a model for preoperative risk assessment. *CHEST Journal*, 102(1):36–44, 1992.

[277] Undefined. Cardiac output as a function of right atrial pressure. http://rfumsphysiology.pbworks.com/f/cc2.bmp.

[278] B. Vallet, C. Chopin, S. E. Curtis, B. A. Dupuis, F. Fourrier, H. Mehdaoui, B. LeRoy, A. Rime, C. Santre, and P. Herbecq. Prognostic value of the dobutamine test in patients with sepsis syndrome and normal lactate values: a prospective, multicenter study. *Critical Care Medicine*, 21(12):1868–1875, Dec 1993.

[279] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[280] Vladimir Vapnik. The support vector method of function estimation. In *Nonlinear Modeling*, pages 55–85. Springer, 1998.

[281] Georg Varga, Jan Ehrchen, Athanasios Tsianakas, Klaus Tenbrock, Anke Rattenholl, Stephan Seeliger, Matthias Mack, Johannes Roth, and Cord Sunderkoetter. Glucocorticoids induce an activated, anti-inflammatory monocyte subset in mice that resembles myeloid-derived suppressor cells. *Journal of leukocyte biology*, 84(3):644–650, 2008.

[282] M. Varpula, M. Tallgren, K. Saukkonen, L. M. Voipio-Pulkki, and V. Pettila. Hemodynamic variables related to outcome in septic shock. *Intensive Care Medicine*, 31(8):1066–1071, 2005.

[283] Jose Vina, Consuelo Borras, and Mari Carmen Gomez-Cabrera. Overweight, obesity, and all-cause mortality. *JAMA*, 309(16):1679–1679, 2013.

[284] J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, and L. G. Thijs. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. on behalf of the working group on sepsis-related problems of the european society of intensive care medicine. *Intensive Care Medicine*, 22 (7):707–710, Jul 1996.

[285] Jean-Louis Vincent and Marjorie Beumier. Diagnostic and prognostic markers in sepsis. *Expert Review of Anti-Infective Therapy*, 11(3):265–275, Mar 2013. doi: 10.1586/eri.13.9. URL http://dx.doi.org/10.1586/eri.13.9.

[286] Jean-Louis Vincent and Herwig Gerlach. Fluid resuscitation in severe sepsis and septic shock: an evidence-based review. *Critical care medicine*, 32(11):S451–S454, 2004.

[287] Jean-Louis Vincent and Rui Moreno. Clinical review: scoring systems in the critically ill. *Critical Care*, 14(2):311, 2010.

[288] Jean-Louis Vincent and Mervyn Singer. Critical care: advances and future perspectives. *The Lancet*, 376(9749):1354–1361, 2010.

[289] Jean-Louis Vincent, Jordi Rello, John Marshall, Eliezer Silva, Antonio Anzueto, Claude D Martin, Rui Moreno, Jeffrey Lipman, Charles Gomersall, Yasser Sakr, Konrad Reinhart, and E. P. I. C. II Group of Investigators. International study of the prevalence and outcomes of infection in intensive care units. *JAMA*, 302(21):2323–2329, Dec 2009.

[290] Jean-Louis Vincent, Steven M Opal, John C Marshall, Kevin J Tracey, et al. Sepsis definitions: time for change. *Lancet*, 381(9868):774–775, 2013.

[291] JL Vincent. *Sepsis and Non-Infectious Systemic Inflammation*. Wiley-vch Verlag GmbH and Co, 2009.

[292] JL Vincent. Definition, monitoring, and management of shock states. *Intensive and Critical Care Medicine*, pages 143–150, 2009.

[293] J.L. Vincent, S.M. Opal, and J.C. Marshall. Ten reasons why we should not use severity scores as entry criteria for clinical trials or in our treatment decisions. *Critical Care Medicine*, 38 (1):283, 2010.

[294] Keith R Walley, Nicholas W Lukacs, Theodore J Standiford, Robert M Strieter, and Steven L Kunkel. Balance of inflammatory cytokines related to severity and mortality of murine sepsis. *Infection and immunity*, 64(11):4733–4738, 1996.

[295] C. Van Walraven, P. C Austin, Alison Jennings, Hude Quan, and Alan J Forster. A modification of the elixhauser comorbidity measures into a point system for hospital death using administrative data. *Medical Care*, 47(6):626–633, Jun 2009. doi: 10.1097/MLR.0b013e31819432e5. URL http://dx.doi.org/10.1097/MLR.0b013e31819432e5.

[296] Jiaqi Wang, Xindong Wu, and Chengqi Zhang. Support vector machines based on k-means clustering for real-time business intelligence systems. *International Journal of Business Intelligence and Data Mining*, 1(1):54–64, 2005.

[297] Peter A. Ward and Markus Bosmann. A historical perspective on sepsis. *American Journal of Pathology*, 181(1):2–7, Jul 2012. doi: 10.1016/j.ajpath.2012.05.003. URL http://dx.doi.org/10.1016/j.ajpath.2012.05.003.

[298] Lorraine B Ware, Tatsuki Koyama, D Dean Billheimer, William Wu, Gordon R Bernard, B Taylor Thompson, Roy G Brower, Theodore J Standiford, Thomas R Martin, and Michael A Matthay. Prognostic and pathogenetic value of combining clinical and biochemical indices in patients with acute lung injury. *CHEST Journal*, 137(2):288–296, 2010.

[299] Curtis H Weiss, Farzad Moazed, Colleen A McEvoy, Benjamin D Singer, Igal Szleifer, Luís AN Amaral, Mary Kwasny, Charles M Watts, Stephen D Persell, David W Baker, et al. Prompting physicians to address a daily checklist and process of care and clinical outcomes: a single-site study. *American Journal of Respiratory and CriticalCare Medicine*, 184(6):680, 2011.

[300] Stacey-Ann Whittaker, Mark E Mikkelsen, David F Gaieski, Sherine Koshy, Craig Kean, Barry D Fuchs, et al. Severe sepsis cohorts derived from claims-based strategies appear to be biased toward a more severely ill patient population. *Critical care medicine*, 2013.

[301] REO Williams, M Patricia Jevons, RA Shooter, CJW Hunter, JA Girling, JD Griffiths, and GW Taylor. Nasal staphylococci and sepsis in hospital patients. *British medical journal*, 2(5153):658, 1959.

[302] Bradford D Winters, Ayse P Gurses, Harold Lehmann, J Bryan Sexton, Carlyle Jai Rampersad, Peter J Pronovost, et al. Clinical review: checklists-translating evidence into practice. *Critical Care*, 13(6):210, 2009.

[303] David Wong. Analysis and augmentation of a multi-parameter monitoring system for early warning of patient deterioration. Master's thesis, Department of engineering, 2007.

[304] David Wong, David A Clifton, and Lionel Tarassenko. Probabilistic detection of vital sign abnormality with gaussian process regression. In *Bioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference on*, pages 187–192. IEEE, 2012.

[305] Kelly A Wood and Derek C Angus. Pharmacoeconomic implications of new therapies in sepsis. *Pharmacoeconomics*, 22(14):895–906, 2004.

[306] Chih-Hung Wu, Gwo-Hshiung Tzeng, Yeong-Jia Goo, and Wen-Chang Fang. A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. *Expert Systems with Applications*, 32(2):397–408, 2007.

[307] H Yang and John Moody. Feature selection based on joint mutual information. In *proceedings of international ICSC symposium on advances in intelligent data analysis*, pages 22–25. Citeseer, 1999.

[308] J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. *Intelligent Systems and Their Applications, IEEE*, 13(2):44–49, 1998.

[309] D. York. Least-squares fitting of a straight line. *Canadian Journal of Physics*, 44(5): 1079–1086, 1966.

[310] CA Zauner, RC Apsner, A Kranz, L Kramer, C Madl, B Schneider, B Schneeweiss, K Ratheiser, F Stockenhuber, and K Lenz. Outcome prediction for patients with cirrhosis of the liver in a medical ICU: a comparison of the APACHE scores and liver-specific scoringsystems. *Intensive care medicine*, 22(6):559–563, 1996.

[311] Jack E Zimmerman and Andrew A Kramer. A model for identifying patients who may not need intensive care unit admission. *Journal of critical care*, 25(2):205–213, 2010.

[312] Jack E Zimmerman, Andrew A Kramer, Douglas S McNair, Fern M Malila, and Violet L Shaffer. Intensive care unit length of stay: Benchmarking based on acute physiology and chronic health evaluation (APACHE) iv*. *Critical care medicine*, 34(10):2517–2529, 2006.

[313] J.E. Zimmerman, A.A. Kramer, D.S. McNair, and F.M. Malila. Acute physiology and chronic health evaluation (APACHE) iv: Hospital mortality assessment for today's critically ill patients. *Critical Care Medicine*, 34(5):1297, 2006. ISSN 0090-3493.

[314] Julie R Zivin, Theodore Gooley, Richard A Zager, and Michael J Ryan. Hypocalcemia: a pervasive metabolic abnormality in the critically ill. *American journal of kidney diseases*, 37 (4):689–698, 2001.

[315] D.J. Zwickl. *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. PhD thesis, The University of Texas at Austin, 2007.

Acronyms

$P+$ Positive Predictivity Value. 49

Se Sensitivity. 49

A-aDO$_2$ Alveolar-arterial oxygen difference. 36

ABP Arterial blood Pressure. 20

AI Artificial Intelligence. 42

AIDS Acute Immune Dysfunction Syndrome. 36

APACHE Acute Physiology and Chronic Health Evaluation. 25, 33, 36

APS Acute Physiology Score. 36

ATP Adenosine TriPhosphate. 9

AUROC Area Under the Curve. i, 27, 28, 31, 67–69

B-cell Bursa of Fabricius cell. 5

BIDMC Beth Israel Deaconess Medical Centre. 18

BMI Body Mass Index. 24

BN Bayesian Network. 31

BR Breathing Rate. 20

BUN Blood Urea Nitrogen. 36

CD Cluster of Differentiation. 7

CDF Cumulative Distribution Functions. 50

CHC Chronic Health Condition. 24, 25

CMO Comfort Measures Only. 17

DNA DeoxyriboNucleic Acid. 7, 55

DNI Do Not Intubate. 17

DNR Do Not Resuscitate. 17

EMR Electronic Medical Record. 16

EPV Events Per Variable. 47

FN False Negative. 27

FP False Positive. 27

GA Genetic Algorithm. i, 55, 56, 63

GCS Glasgow Coma Scale. 31, 36

GLM Generalized Linear Model. 50

GPS Global Postionning System. 44

HDFR Hypotension Despite Fluid Resuscitation. 2

HIF Hypoxia-Inducible Factor. 9

HIS Hospital Information System. 18, 19

HL-gof Hosmer-Lemeshow goodness-of-fit test. 27, 28

HR Heart Rate. 20

ICD International Classification Disease. 24, 31, 35

ICU Intensive Care Unit. i, 1, 2, 6, 17–20, 24, 25, 36

IDI Integrated Discriminative Improvement. 27, 29

If InterFeron. 8, 9

IL InterLeukin. 8, 9

LASSO Least Absolute Shrinkage And Selection Operator. 48, 68

LOS Length Of Stay. 16, 36

LPB LipoPolysaccharide-Binding protein. 7

LPS LipoPolySaccharide. 7

LR Logistic Regression. 31, 45

LSE Least Squares estimate. 47, 48

LTA LipoTeichoic Acid. 7

Machine Learning Machine Learning. 42, 43, 45

MAP Mean Arterial Pressure. 36

MD Lymphocyte antigen 96. 7

MIMIC-II Multi-parameter Intelligent Monitoring in Intensive Care II. i, 18, 19, 25, 31, 36, 48, 67

MODS Multiple Organ Dysfunction Syndrome. 6, 16

MSE Multi-Scale Entropy. 65

Na Sodium. 36

NLL Negative Log Likelihood. 49

NN Neural Network. 31

NO Nitric Oxide. 8, 9

NRI Net Reclassification Improvement. 27, 29

PaO_2 Partial Arterial pressure of O_2. 36

PaO_2/FiO_2 Ratio of partial pressure of arterial O_2 to the fraction of inspired O_2. 31

Pattern Recognition Pattern Recognition. 42, 43

PG PeptidoGlycan. 7

RBF Radial Basis Function. 46, 62, 63

ROC Receiver Operating Characteristic. 27

RT Regression Tree. 31

SA Simulated Annealing. 63

SAPS Simplified Acute Physiology Score. i, 2, 33, 35

SE Standard Error. 27, 28

SEV Systolic Ejection Volume. 52

SIRS Systemic Inflammatory Response Syndrome. 2, 6, 24

SOFA Sepsis-related Organ Failure Assessment. 31

SpO_2 Arterial Oxygen Saturation. 20

SQL Structured Query Language. 19

SSE Sum Square Error. 47

SVM Support Vector Machine. 46, 50, 62, 63

T-cell Thymus cell. 5

TAT Turnaround Time. 6

TLR Toll-Like Receptor. 7

TN True Negative. 27

TNF Tumour Necrosis Factor. 8, 9

TP True Positive. 27

WBC White Blood Cell. 4, 36

Appendix A

Description of the variables

Table A.1: *Description of tables and item IDs used for the extraction of treatments and intervention in the Multi-parameter Intelligent Monitoring in the Intensive Care II (MIMIC-II) database.*

	Fluid extraction details
Table:	ioevents
Unique ID:	icustay ID
Value field:	volume
Unit Field:	volumeuom
Time field:	charttime
	Item IDs
Coloids	232, 233, 246, 258, 342, 548, 662, 33, 1894, 1438, 2550, 4509, 4588, 3610, 2688, 2832, 2944, 2975, 4135, 6257, 6313, 5410, 5487, 623, 2551, 2731, 5019, 2349, 6564, 6729, 3237, 3317, 3353, 4203, 6170, 5152, 5403, 5426
Ringers	204 , 219 , 295 , 324 , 518 , 106 , 1030 , 1225 , 1322 , 2350 , 4521 , 4571 , 4654 , 4915 , 4815 , 3672 , 2600 , 2615 , 2759 , 2978 , 6525 , 4110 , 4184 , 5102 , 6188 , 5229 , 5284 , 5532 , 5642 , 6781 , 753 , 5859 , 4399 , 4452 , 6764 , 142 , 1819 , 1452 , 4099 , 5548
Saline	194, 249, 423, 432, 438, 444, 448, 480, 512, 568, 631, 776, 811, 818, 850, 865, 1008, 1051, 1132, 1204, 1215, 1237, 1245, 1291, 1356, 1371, 1380, 1392, 1428, 1432, 1467, 1490, 1491, 1526, 1540, 1543, 1648, 1674, 1695, 1719, 1731, 1754, 1790, 1878, 1896, 1977, 2019, 2039, 2082, 2087, 2192, 2231, 2242, 2271, 2297, 2363, 2401, 2453, 2532, 2548, 2606, 2639, 2715, 2749, 2820, 2844, 3186, 3718, 3750, 3759, 3849, 3986, 4053, 4126, 4160, 4171, 4200, 4263, 4290, 4426, 4431, 4440, 4491, 4533, 4633, 4735, 4741, 4858, 4894, 4983, 5073, 5079, 5116, 5278, 5341, 5480, 5557, 5561, 5989, 6314, 6376, 6400, 6404, 6782, 6800
D5 Saline	214, 216, 297, 629, 706, 984, 1024, 1959, 1969, 1975, 1317, 1320, 1469, 5062, 5098, 3784, 4321, 4320, 5374, 4829, 3736, 154, 1604, 1826, 1935, 883, 299, 615, 4248, 5212

Hypotonic	389, 542, 695, 781, 151, 180, 854, 936, 1230, 1261, 4496, 4792, 5975, 6584, 4239, 4439, 2290, 2354, 943, 1079, 1096, 1107, 1260, 5615

Medication extraction details

Table:	medevents
Unique ID:	icustay ID
Value field:	dose
Unit Field:	doseuom
Time field:	realtime

	Item IDs
Insuline	45, 100, 310
Pressors	42, 43, 44, 46, 47, 51, 119, 100, 120, 125, 127, 128, 306, 307, 309
Sedatives	124, 141, 118, 149, 150, 308, 163, 131, 167

Procedures extraction details

Table:	procedureevents
Unique ID:	subject ID
Value field:	sequence_num
Unit Field:	sequence_num
Time field:	proc_dt

	Item IDs
MechVent	101729, 101781, 101782, 101783
RTT	100576, 100586, 100593, 100594, 100622, 100977
Graft	101647, 101648, 101649, 101700, 100458, 100159, 100161, 100462, 100463,100464, 100470
Bypass	100461, 100462, 100463, 100464
Intub	1100323, 100339, 100341, 101749, 101750, 101776, 101794, 101750

Abbreviations: MechVent, Mechanical ventilation; RRT, Renal replacement therapy; Bypass, Bypass surgery; Intub, Inutbation.

Table A.2: *This tab describes variables and define their abbreviation. Variables are recorded at the time of admission (first six lines) or defined as the median value of 2-hours-long or 24-hours-long time window. For the latter, variables names have been given the prefix DAY-. Variables are preceded in this report by one of the following prefixes: PRE-, PER- and POST- indicating if the measurement has been taken before, during or after the hypotensive event, respectively.*

Variables	Description	Unit
AGE-Y	Age at the time of admission.	year
WEIGHT-FIRST	Weight at the time of admission.	kg
HEIGHT	Height at the time of admission.	cm
HE-LENGTH	Length of the hypotensive event	days
HE-ONSET-HOURS	Time from admission to onset of hypotensive event.	hours
SAPSI-FIRST	Automatically calculated Simplified Acute Physiology Score (SAPS)-I score. Please refer to [97] for details.	-
Van Walraven	Pre-computed score based on a 30 comorbidities recommanded by [83]. Please refer to [295] for details.	-
HR	Heart Rate.	Beat per minute (bpm)
ABPS	Systolic Arterial Blood Pressure ($ABP_{Systolic}$) is the value of arterial blood pressure at systole.	mmHg
ABPD	Diastolic Arterial Blood Pressure ($ABP_{Diastolic}$) is the value of arterial blood pressure at diastole.	mmHg
ABPM	Mean Arterial Blood Pressure (ABP_{Mean}) is defined by $2/3 * ABP_{sys} + 1/3 * ABP_{dia}$.	mmHg
BR	Breathing Rate (BR).	Breaths per minute (bpm)
SPO2	Arterial Oxygen Saturation (SpO_2).	%
CVP	Central Venous Pressure (CVP).	mmHg
DAY-RESP	Ratio of partial pressure of arterial O_2 to the fraction of inspired O_2 (PaO_2/FiO_2) gives information on the respiratory function.	mmHg/torr
DAY-GCS	Glasgow Coma Scale (GCS).	-
DAY-TEMP	Temperature.	Celcius
DAY-ARTPH	Arterial pH.	-
DAY-BICARB	Bicarbonate (HCO_3^-) acts as an hydrogene buffer in the blood which maintains a constant pH.	mg/dL
DAY-BUN	Blood Urea Nitrogen (BUN) is a measure of nitrogen in the blood in the form of urea that gives information about the renal function.	mg/dL
DAY-HEMATOCRIT	Hematocrit is the proportion of blood volume occupied by red blood cells. Red cells are involved in oxygene transportation.	%
DAY-HBG	Haemoglobin blood level. Haemoglobin is a protein involved in oxygene transportation.	mg/L
DAY-PLAT	Platelets blood level. Platelets are involved in the cloating process.	g/L
DAY-CA	Calcium blood level.	mg/dL
DAY-CL	Chlorite blood level.	mg/dL
DAY-CREAT	Creatinine bood level. Creatinine is produced at a constant rate in the body : its increase usually indicates a damaged clearance mechanism which relates kidney function	μg/dL
DAY-GLUC	Glucose blood level.	mg/dL
DAY-LACTAC	Lactate blood level. Lactate is a by-product of anaerobic metabolism indicative of lack of adequate perfusion.	mg/dL
DAY-MG	Magnesium blood level.	mg/dL
DAY-P	Phosphorus blood level.	mg/dL
DAY-K	Potatium blood level.	mg/dL
DAY-NA	Sodium blood level.	mg/dL
DAY-WBC	White Blood Cell (WBC) cells in plasma. White cells or leukocytes are involved in the immune system response to danger signals.	-
DAY-BILI	Bilirubin blood level. Bilirubin is the yellow product of heme catabolism that is mainly found in hemoglobin and which blood level is indicating of liver function.	mg/dL
DAY-ALT	ALanine aminoTransferase (ALT) blood level related to liver function.	IU/L
DAY-INR	International Normalized Ratio (INR) is related to coagulation pathways.	-
DAY-PACO2	Partial CO_2 arterial blood pressure.	%
TOTALFLUIDVOL	Total amount of fluid given to patient during HE LENGTH.	mL
FLUIDRATE	TOTALFLUIDVOL/HE LENGTH is the fluid rate of resuscitation during hypotensive event.	mL/hours
PRESSORS	Number of different pressors given to patient during the hypotensive event.	-

Table A.3: *Description of item IDs used for the extraction of different variables*

	Table description		
	Table:	chartevents	labevents
	Unique ID:	icustay_id	icustay_id
	Item field:	itemid	itemid
	Value field:	value1num	valuenum
	Unit Field:	value1uom	valueuom
	Time field:	charttime	charttime
Description	Label	Item IDs	
Temperature	Temp	676, 677	
Diastolic BP (mmHg)	NIDiasABP	6, 455 [a]	
Systolic BP (mmHg)	NISysABP	6, 455	
Mean BP (mmHg)	NIMAP	456, 1149,751	
Invasive Systolic BP (mmHg)	SysABP	51	
Invasive Diastolic BP (mmHg)	DiasABP	51 [a]	
Invasive Mean BP (mmHg)	MAP	52	
Heart Rate (/minute)	HR	211	
Respiration Rate (/minute)	RespRate	615,618	
FIO2 (%)	FiO2	3420, 190, 3422, 3421	
PH (-)	pH	780, 1126, 4753, 3839, 4202, 865	50018
Partial Arterial pressure of O_2 (PaO_2) (mmHg)	PaO2	4203, 3785, 3837, 3838	50019
Partial Arterial pressure of CO_2 ($PaCO_2$) (mmHg)	PaCO2	4201, 3784, 3835, 3836	50016
Sodium (mmEq)	Na	837, 4195, 3726, 3803, 1536,	50159
Glucose (mg/dL)	Glucose	1455, 807, 811, 3447, 1310, 3744, 3745, 1529, 1812, 2338, 2416	50006
Creatinine (mmEq)	Creatinine	791, 3750,1525	50090
BUN (mmEq)	BUN	781, 1162, 5876, 3737	
Albumin (mmEq)	Albumin	772, 3066, 3727, 1521, 2358	50060
Bilirubin (mmEq)	Bilirubin	5032, 5045, 4354, 5483, 5543, 4948, 848, 1538	50170
Haematocrit (HCT) (%)	HCT	813	50383
WBC (-)	WBC	861, 1127, 4200, 1542	50468
INR (-)	INR	815, 1530	50399
Potassium (mmEq)	K	1535, 829	50149
Phosphorous (mmEq)	P		50148
Magnesium (mmEq)	Mg	821, 1532	50140
Lactate (mmEq)	Lactate	818, 1531	50010
Chloride (mmEq)	Chloride	788, 4193, 3747, 1523	50083
Calcium (mmEq)	Calcium	1522, 3746, 786	50079
Platelets (-)	Platelets	828, 3789, 6256	
Haemoglobin	Haemoglobin	814	
Bicarbonates (mmEq)	HCO_3	787, 3810	
GCS (-)	GCS	198	
Eye Open (-)	EyeOpen	184 [b]	
Verbal response (-)	VerbalResp	723 [b]	
Motor response (-)	MotorResp	454 [b]	
SAPS-I (-)	SAPSI	20001	
Respiratory Sepsis-related Organ Failure Assessment (SOFA) (-)	RespSOFA	20002	
Hepatic SOFA (-)	HepSOFA	20003	
Hemato SOFA (-)	HemSOFA	20004	
Press Cardio SOFA(-)	PrCardSOFA	20005	
MAP Cardio SOFA (-)	MaCardioSOFA	20006	
Neuro SOFA (-)	NeuroSOFA	20007	
Cardiac Output (L/min)	CO	90, 89, 1601, 2112	
CVP (mmHg)	CVP	113, 3345, 1103	
SpO_2 (%)	SpO2	646, 6719	
Arterial Oxygenation Saturation (SaO_2) (%)	SaO2	834,3609,4833,3495	
Alkaline phosphatase (mmEq)	ALP	3728	50061
Alanine transaminase (mmEq)	ALT	769	50062
Aspartate transaminase (mmEq)	AST		50073
Cholesterol (mmEq)	Cholesterol	789,3748,1524	50085
Troponin-I (mmEq)	TropI		50188
Troponin-T (mmEq)	TropT		50189
Rikers' scale (-)	RikersScale	1337 [b]	
MechVent (-)	MechVent	720, 721, 722	

[a] indicates where value was taken from $value2num$ instead of $value1num$,
[b] indicates that the value was taken from $value1$ attribute (text value).

Table A.4: *Description of the International Classification of Diseases (ICD)-9 codes to extract some Chronic Health Condition (CHC). Because codes are organized in a tree structure, a '%' symbol (following the oracle syntax) indicates that all the sub-codes are included.*

	Table description	
	Table:	ICD9
	Unique ID:	hadm_id
	Item field:	CODE
	Value field:	sequence
Description	Label	Item IDs
Hepatic Failure	HepFail	570%[a]
AIDS	AIDS	042%
Cirrhosis	Cirrhosis	571%
Myeloma	Myeloma	203.0%
Metcan	Metcan	196.%, 197.%, 198.%, 199.%, 200.%, 201.%, 202.%, 203.%, 205.%, 208.%

[a] the '%' indicates that zero or more characters can come in place.

Table A.5: *Description of parameter used for the extraction of microbiology events. From the* $microbiologyevents$ *table, three fields are used to extract the organism, the site of sample collection, and the antibiotic test interpretation: org_itemid, spec_itemid, and interpretation, respectively.*

Table description		
	Table:	Microbiologyevents
	Unique ID:	hadm_id
	Item field:	org_itemid
	Value field:	interpretation
Description	Label	Item IDs
Staph. Aur. Coag.	Org_StaphAurCoag	80023
Enteroco. SP	Org_EnterocoSP	80053
Staph. Coag.	Org_StaphCoag	80155
E. Coli.	Org_Ecoli	80002
Pseudo. Monas. Aeru.	Org_PseudomonasAeru	80026
Klemsiel. Pneumo.	Org_KlemsielPneumo	80004
Enteroco. Faecium	Org_EnterocoFaecium	80168
Entero. Bact. Cloac.	Org_EnteroBactCloac	80008

Table description		
	Table:	Microbiologyevents
	Unique ID:	hadm_id
	Item field:	spec_itemid
	Value field:	interpretation
Description	Label	Item IDs
Sputum	Site_sputum	70062
Swab	Site_swab	70040, 70067, 70068, 70069, 70070, 70071, 70084, 70092
Urine	Site_urine	70077, 70078, 70079, 70080, 70081, 70082
Catheter	Site_iv	70023
Blood	Site_blood	70007, 70011, 70012, 70012, 70013, 70014, 70015, 70016, 70017, 70049, 70052