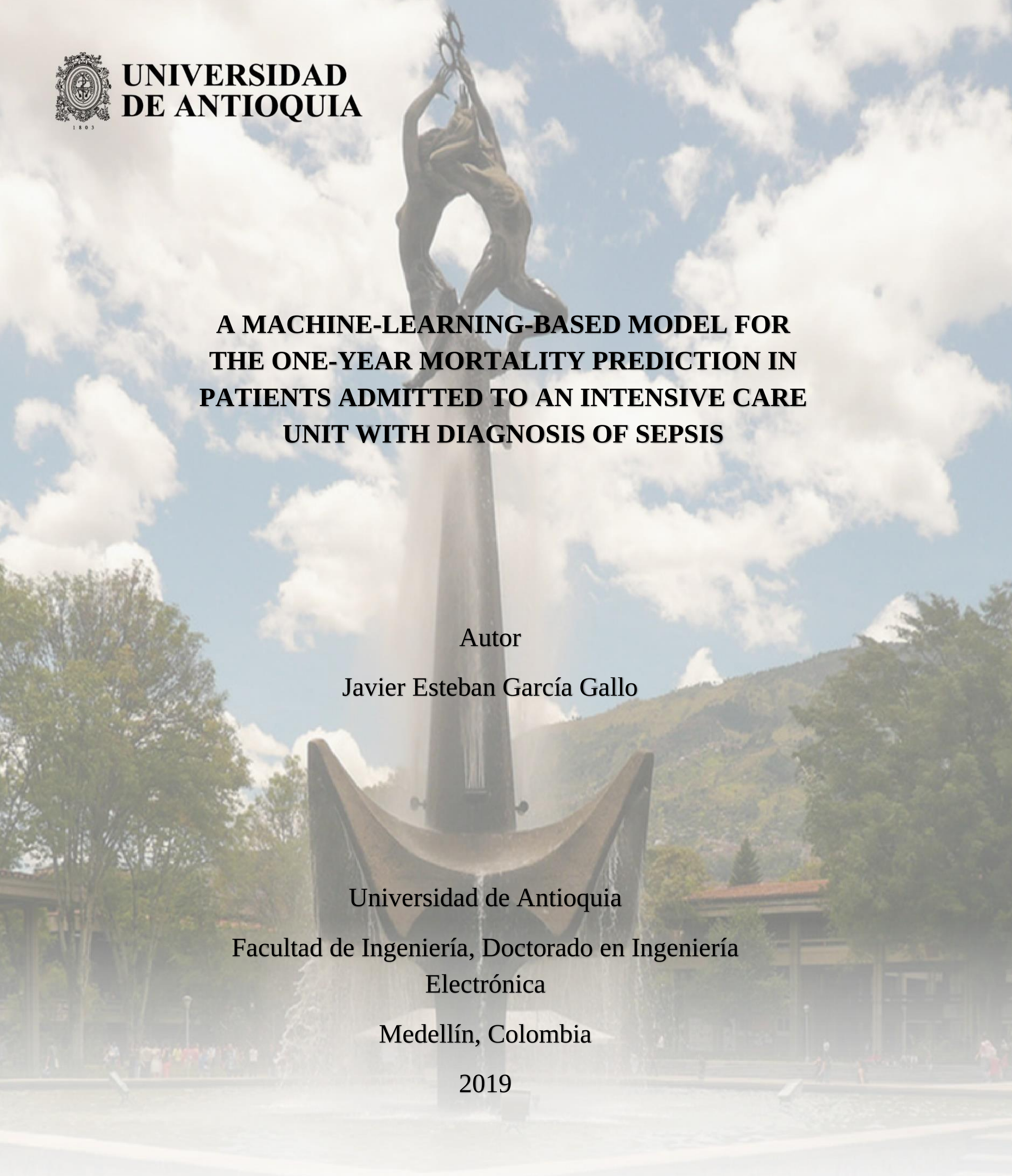


UNIVERSIDAD
DE ANTIOQUIA

A MACHINE-LEARNING-BASED MODEL FOR THE ONE-YEAR MORTALITY PREDICTION IN PATIENTS ADMITTED TO AN INTENSIVE CARE UNIT WITH DIAGNOSIS OF SEPSIS

Autor

Javier Esteban García Gallo

Universidad de Antioquia

Facultad de Ingeniería, Doctorado en Ingeniería Electrónica

Medellín, Colombia

2019

A Machine-learning-based Model for the One-year Mortality Prediction in Patients Admitted to an Intensive Care Unit with Diagnosis of Sepsis

Javier Esteban García-Gallo

Intelligent Information Systems Lab (In2Lab),
Universidad de Antioquia UdeA, Medellín, Colombia

Tesis como requisito para optar al título de:
Doctor en Ingeniería Electrónica.

Director:

John Freddy Duitama Muñoz, PhD

Profesor Asociado

Intelligent Information Systems Lab (In2Lab),
Universidad de Antioquia UdeA, Medellín, Colombia

Asesor:

Nelson Javier Fonseca-Ruiz MD

Universidad de Antioquia

Facultad de Ingeniería, Doctorado en Ingeniería Electrónica

Medellín, Colombia

2019

A mis padres, por su entrega y cariño

A mi amada esposa, por su apoyo y paciencia

A mis hermosos hijos, quienes me exhortan a ser mejor cada día

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Professor John Freddy Duitama Muñoz for the continuous support of my PhD study and research, for his trust, motivation and for sharing his vast experience with me. I could not have imagined having a better mentor.

I would like to thank the rest of my thesis advisors: Dr. Nelson Javier Fonseca Ruiz, Dr. Leo Anthony Celi, for their inspiration, insightful comments, and willingness to answer my questions.

My sincere thanks also goes to Professor Nassir Navab, Dr. Shadi Albarqouni, and PhD student Anees Kazi, for accepting me for the internship in their group and make me feel welcome and appreciated, his advice during my stay in their laboratory made this work much better.

I want to express my gratitude to Universidad de Antioquia for educating me both academically and personally and to Departamento Administrativo de Ciencia, Tecnología e Innovación (Colciencias) for the support.

Finalmente, pero no menos importante, agradezco a mi familia. A mi madre Martha Cecilia García Gallo, por su esmero y cariño. A mi padre Luis Javier Garcia López, por enseñarme la importancia de una ética implacable. A mi amada esposa Sara Isabel Duque, por brindarme los momentos más felices de mi vida. A mis hijos por su alegría y motivación.

Javier Esteban García Gallo.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The Intensive Care Unit (ICU) is a hospital department that provides intensive treatment to patients with severe and life-threatening conditions. The primary function of the ICU is to deliver care which cannot be administered in other areas of the hospital. Patients in the ICU are the most heavily monitored patients in the entire hospital; for this reasons the ICU is a data rich environment, even to the point of exhaustion.

The vast amount of data obtained from a single patient in an intensive care unit makes it humanly impossible to organize and interpret it in the required time, thus, scores that model the patient severity and can be related with the mortality have been created. The primary motivation of this scores was to derive further insight of the patient condition and improve patient care.

Traditionally, this scores are population-based and provide statistically rigorous results for an average patient, and are useful to guide prognostication, to assess ongoing disease development and organ function, to compare ICU performance over time and across units and to compare clinical trial population outcomes but, pitifully, they are not precise enough to draw conclusions about groups of patients that share a relevant clinical condition, like a particular disease, and even less to be used for individual prediction of outcomes.

When standard scores do not fit the data of a specific population well enough, two approaches to adapting them for use among patients with the specific condition have been used. One approach would be to modify the traditional score by adapting them for use specifically among patients that share a condition, which we will be referring as adjusted models. The other approach would be to develop entirely new models based on a population that shares a common characteristics and that incorporates additional variables that could potentially enhance accuracy, which we will be referring as customized models.

Sepsis patients are a specific population that is especially vulnerable, since they present a high in-hospital mortality of 25–30% and patients with sepsis are frequently cared for in ICUs, either because sepsis itself led to their admission or because sepsis developed as a complication of their admission for other reasons; moreover, it has been reported that sepsis survivors had substantially increased risks of all-cause mortality, as well as major health complications at 1 year after discharge when compared with the general population.

For sepsis patients within the ICU, mortality prediction has been accessed through both adjusted and customized models; however, approaches addressed so far have focused on the in-hospital mortality prediction, and no methods have been proposed to identify and predict long-term risk and mortality in sepsis patients that are being taken care of in the ICU.

According to the above, in this work, we present the development of a model that goes beyond the prediction of in-hospital mortality and alert those patients who may have a poor prognosis after being discharged from the hospital, and we formulate our research question as follows: Among adult ICU patients, is it possible to identify those who are at risk of dying one year after their sepsis related admission using demographic variables, comorbidities and physiological data obtained during the first 24 hours of their ICU stay?

In order to answer such a question, we used three approaches. First we developed a custom one-year mortality prediction model using a Stochastic Gradient Boosting (SGB) technique. The model was based on the data of 5650 ICU patient's admissions that were retrospectively identified as having sepsis, and used 132 predictors, obtained from variables found in the literature review or suggested by experts. In the first approach, we also used two techniques to measure the importance of the used predictors, and we found 17 predictors that allowed us to develop an SGB model with a performance similar to the complete model (which uses all the 132 predictors).

In the second approach, we developed a methodology that allows the stratification of patients according to their one-year mortality risk. For this, we extended our study cohort using two additional retrospective criteria for sepsis identification and focusing only on the variables that were relevant (according to the results of the previous approach) or that were routinely taken to patients within the ICU, obtaining 15082 admissions; From said cohort we developed two scores systems that are correlated with the one-year mortality risk of the patients.

Although the developed customized models for sepsis patient within the ICU proved far outperform adjusted scores for the one-year mortality prediction task, they continue to be population-based and therefore they provide "the average best choice" for sepsis patients. For this reason, in the third approach, we also propose and evaluate the generation of personalized models based on patient similarity metrics. The goal of this personalized models is to identify patients who are similar to a new patient and derive insights from the data of those similar patients to provide personalized predictions.

Personalized models has been widely used for predictions in several fields, including music, movies and e-commerce, however, there are still very few studies that focus on personalized prediction models based on health data prediction. Moreover, no studies have been reported in which personalized models are developed from a population known to be very homogenous, such as our study population, where it is known that all patients have infection, organ dysfunction, and ICU stays of more than 24 hours.

The developed models, with the three approaches, showed discrimination superior to adjusted models based on traditional severity scores and, the population based methodologies also presented adequate calibration. Specifically, our personalized models demonstrated the value of patient similarity metrics in outcome prediction modeling and showed superiority when compared to population-based models. Also, since we focused on long-term mortality prediction, these models successfully identify those patients who are at risk of dying one year after their sepsis related admission using demographic variables, comorbidities and physiological data obtained during the first 24 hours of their ICU stay, indicating early, which patients should be accompanied, observed attentively and provided with additional care that improve their quality of life.

Finally, in order to enable the clinical use of the machine learning models developed for the prediction of one-year mortality of sepsis patients within the ICU, we developed a software based on the models that presented a better performance and the functionalities that are considered useful so that intensivist can obtain details of the particular condition of each patient and provide better care.

PART 1: CLINICAL SCENARIO

This part of the thesis sets the stage for outcome prediction within the Intensive Care Units, details the concept of sepsis, raises our research question, and outlines the methodology to answer it. Chapter 1 presents the way in which mortality are currently being predicted within the ICU. Chapter 2 contextualizes the concept of sepsis and indicates why it is a condition that is worth studying and Chapter 3 presents how the general perspective detailed in the first two chapters is transformed in a pertinent research question, it also includes our study design and the selected outcome of interest.

CHAPTER 1. INTENSIVE CARE UNITS

1.1 Introduction

An Intensive Care Unit (ICU) is a special area of a hospital that provides intensive treatment to patients with severe and life-threatening conditions; these patients are constantly monitored and are cared by highly trained personnel. A large number of physiological and laboratory variables are gathered daily from the patients in an ICU, which allows caregivers to track their progress.

However, the vast amount of data obtained from a single patient in an intensive care unit makes it humanly impossible to organize and interpret it in the required time [1]; for this reason, different types of indicators that seeks to summarize the patient's condition have been developed.

The indicators used in medical practice within the ICU can be broadly divided into: those that synthesize multiple physiological and demographic data into a single number that represents the severity of the illness of a patient and those based on a single physiological measure, also known as a biomarker, which is used for interpretation, evaluation and understanding of different disease processes.

The indicators of the first group are developed from statistical analysis of the data collected for a large number of patients and seek to express in a single number the severity of a patient's illness; in general this score increases with the mortality risk. These kind of classification systems are used to determine the risk in population studies conducted in ICU, and provide a method for benchmarking between intensive care units of different hospitals [2–6].

The indicators of the second group are proposed when a physiological measurement can be used to differentiate normal biological processes from pathological ones or to indicate the response to a therapeutic intervention. They are based on a deep understanding of the causes that vary such measurement within the organism, and are corroborated by epidemiological studies. The main uses of these kinds of indicators are predicting prognosis and guide treatment of patients.

In this work, we present the development of a composite set of models for the prediction of long-term mortality in sepsis patients within the ICU; These models are based on the methodologies used by the indicators of the first group, but are complemented by physiological measures that have proven their usefulness as sepsis biomarkers.

1.2 Severity-of-illness scoring systems

Scoring systems used in critically ill patients can be broadly divided into scores that assess disease severity and use it to predict outcome and scores that assess the presence and severity of organ dysfunction, in this section we review the most commonly used severity-of-illness scoring systems in each of these two groups. We present different versions of the scores that have been updated over time, and list the variables of the most used version of each of the reviewed scores [7].

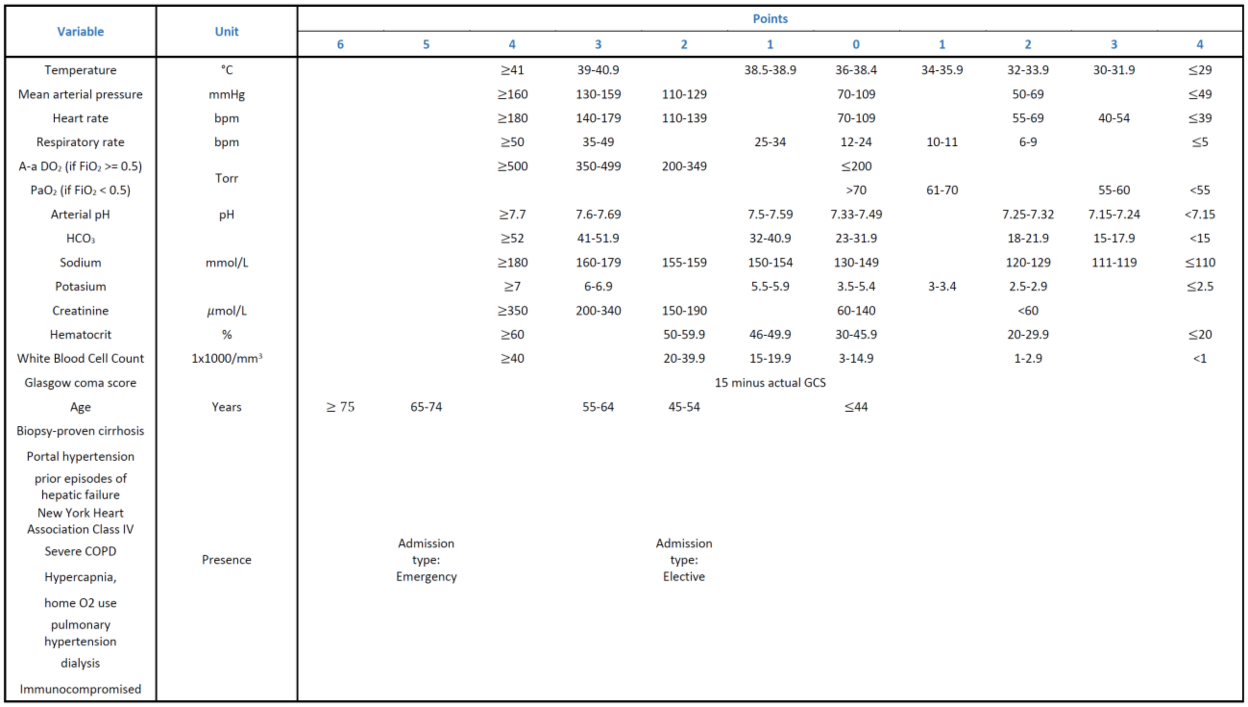
1.2.1 Outcome prediction scores

The outcome prediction scoring systems were developed to indicate the mortality risk of groups of ICU patients, they were not designed for individual prognostication, and usually comprises of two parts, 1) a number assigned to disease severity, commonly known as the score, and a model that gives the probability of hospital mortality of the patients[7, 8] . The following scores are currently used for assessing the acuity of a general ICU population.

1.2.1.1 Acute Physiology and Chronic Health Evaluation

The original Acute Physiology and Chronic Health Evaluation (APACHE) was developed in 1981 to classify groups of patients on the basis of severity of illness. APACHE uses a logistic regression model with hospital mortality as the outcome variable and a set of predictors including comorbidities, age, gender and 34 physiological measures. APACHE contains two parts: a) a physiology score representing the degree of acute illness and b) a preadmission health evaluation indicating health status before acute illness [9]. Four years later appears APACHE II, a simplified version of the previous version, which aims to improve its clinical acceptability. It uses a point score based upon values of 12 routine physiologic measurements (taken during the first 24 h after admission), age and previous health status to provide a general measure of severity of disease. Table 1.1 presents the APACHE II scoring system.

Table 1.1 Acute Physiology and Chronic Health Evaluation Score II. The points presented in the table are the values that are summed to the score when a patient is in a particular group for each of the predictors; for instance, in the temperature of the patient is above 41°C, a four is summed to the total score.



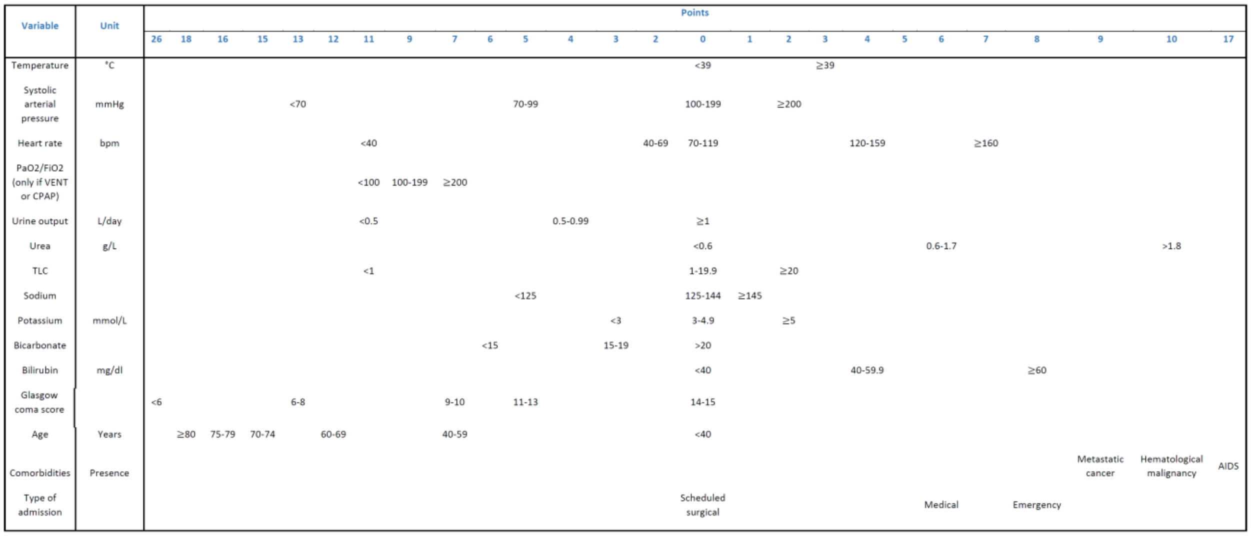
Variable	Unit	Points										
		6	5	4	3	2	1	0	1	2	3	4
Temperature	°C			≥41	39-40.9		38.5-38.9	36-38.4	34-35.9	32-33.9	30-31.9	≤29
Mean arterial pressure	mmHg			≥160	130-159	110-129		70-109		50-69		≤49
Heart rate	bpm			≥180	140-179	110-139		70-109		55-69	40-54	≤39
Respiratory rate	bpm			≥50	35-49		25-34	12-24	10-11	6-9		≤5
A-a DO_2 (if FiO_2 >= 0.5)	Torr			≥500	350-499	200-349		≤200				
PaO_2 (if FiO_2 < 0.5)								>70	61-70		55-60	<55
Arterial pH	pH			≥7.7	7.6-7.69		7.5-7.59	7.33-7.49		7.25-7.32	7.15-7.24	<7.15
HCO_3				≥52	41-51.9		32-40.9	23-31.9		18-21.9	15-17.9	<15
Sodium	mmol/L			≥180	160-179	155-159	150-154	130-149		120-129	111-119	≤110
Potasium				≥7	6-6.9		5.5-5.9	3.5-5.4	3-3.4	2.5-2.9		≤2.5
Creatinine	µmol/L			≥350	200-340	150-190		60-140		<60		
Hematocrit	%			≥60		50-59.9	46-49.9	30-45.9		20-29.9		≤20
White Blood Cell Count	1x1000/mm^3			≥40		20-39.9	15-19.9	3-14.9		1-2.9		<1
Glasgow coma score							15 minus actual GCS					
Age	Years	≥ 75	65-74		55-64	45-54		≤44				
Biopsy-proven cirrhosis	Presence		Admission type: Emergency			Admission type: Elective						
Portal hypertension												
prior episodes of hepatic failure												
New York Heart Association Class IV												
Severe COPD												
Hypercapnia,												
home O2 use												
pulmonary hypertension												
dialysis												
Immunocompromised												

APACHE III appears in 1991 and largely uses the same variables as APACHE II; however, it uses a different way to collect the neurological data -no longer using the Glasgow Coma Scale (GCS), and also adds particularly two important variables: The patient's origin and the lead-time bias. Most recently, APACHE IV was developed using a database of over 100,000 patients admitted to 104 ICUs in 45 hospitals in the United States between 2002 and 2003. In APACHE IV predictor variables are similar to those in APACHE II, but it includes new variables such as urine output, blood urea nitrogen, albumin, bilirubin and glucose; Chronic health conditions (lymphoma, leukemia and metastatic tumor) treatments (Thrombolytic therapy, Mechanical ventilation) and administrative information (ICU admission diagnosis, ICU admission source and Length of stay before ICU admission) were also added. [7, 8, 10].

1.2.1.2 Simplified Acute Physiology Score

The Simplified Acute Physiology Score (SAPS) was developed in 1984, and was intended as a simplification for the original APACHE that differs from APACHE II in the number of variables. SAPS reduces the number of the physiological parameters to 13 and introduces age as new parameter. Later, in 1993, SAPS II was developed and includes 17 variables: 12 physiological variables, age, type of admission, and 3 variables related to underlying disease. The SAPS II score was validated using data from consecutive admissions to 137 ICUs in 12 countries. Table 1.2 presents the SAPS II scoring system [5, 7].

Table 1.2 Simplified Acute Physiology Score II.



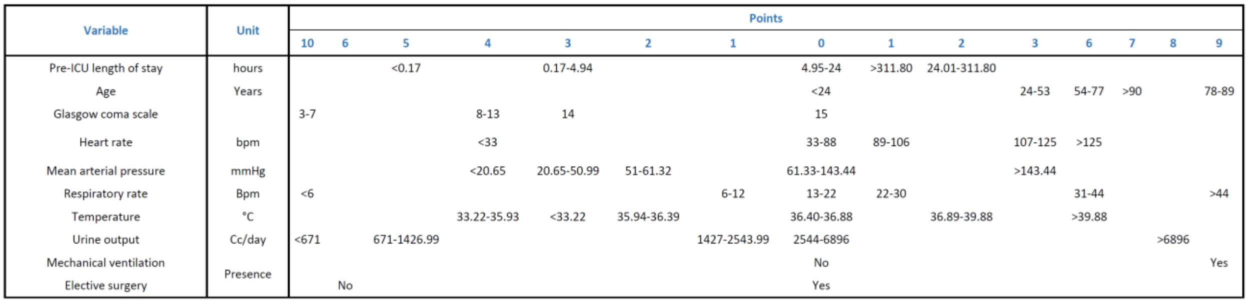
Variable	Unit	Points																									
		26	18	16	15	13	12	11	9	7	6	5	4	3	2	0	1	2	3	4	5	6	7	8	9	10	17
Temperature	°C															<39			≥39								
Systolic arterial pressure	mmHg					<70						70-99				100-199		≥200									
Heart rate	bpm							<40							40-69	70-119				120-159			≥160				
PaO2/FiO2 (only if VENT or CPAP)								<100	100-199	≥200																	
Urine output	L/day							<0.5					0.5-0.99			≥1											
Urea	g/L															<0.6						0.6-1.7				>1.8	
TLC								<1								1-19.9		≥20									
Sodium												<125				125-144	≥145										
Potassium	mmol/L													<3		3-4.9		≥5									
Bicarbonate											<15			15-19		>20											
Bilirubin	mg/dl															<40				40-59.9				≥60			
Glasgow coma score		<6				6-8				9-10		11-13				14-15											
Age	Years		≥80	75-79	70-74		60-69			40-59						<40											
Comorbidities	Presence																								Metastatic cancer	Hematological malignancy	AIDS
Type of admission																Scheduled surgical						Medical		Emergency			

While SAPS II was developed on ICUs in Western Europe, SAPS III included a world-wide database of 16,784 patients; additionally, SAPS III includes 20 variables divided into three parts related to patient characteristics prior to admission, the circumstance of the admission, and the degree of physiological derangement within 1 hour before or after ICU admission. Unlike the other scores, SAPS III includes customized equations for prediction of hospital mortality in seven geographical regions: Australasia; Central, South America; Central, Western Europe; Eastern Europe; North Europe; Southern Europe, Mediterranean; and North America [7, 8].

1.2.1.3 Oxford Acute Severity of Illness Score

The Oxford Acute Severity of Illness Score (OASIS) was developed in 2013. It has equivalent discrimination and calibration of the APACHE IV from which it was derived. OASIS score uses the worst measurements from the first 24 hours of ICU admission [11], and uses 81,000 admissions from a large multi-center database collected by the Cerner corporation (Kansas City, MO, United States). To select subsets of the available features OASIS uses a genetic algorithm, and uses a customized Particle Swarm Optimization method to calculate the score. As consequence, two logistic regressions using OASIS as a covariate were developed for both ICU and hospital mortality[12]. Table 1.3 presents the OASIS scoring system.

Table 1.3. Oxford Acute Severity of Illness Score



Variable	Unit	Points														
		10	6	5	4	3	2	1	0	1	2	3	6	7	8	9
Pre-ICU length of stay	hours			<0.17		0.17-4.94			4.95-24	>311.80	24.01-311.80					
Age	Years								<24			24-53	54-77	>90		78-89
Glasgow coma scale		3-7			8-13	14			15							
Heart rate	bpm				<33				33-88	89-106		107-125	>125			
Mean arterial pressure	mmHg				<20.65	20.65-50.99	51-61.32		61.33-143.44			>143.44				
Respiratory rate	Bpm	<6						6-12	13-22	22-30			31-44			>44
Temperature	°C				33.22-35.93	<33.22	35.94-36.39		36.40-36.88		36.89-39.88		>39.88			
Urine output	Cc/day	<671		671-1426.99				1427-2543.99	2544-6896						>6896	
Mechanical ventilation	Presence								No							Yes
Elective surgery			No						Yes							

1.2.1.4 Severity score comparisons

Table 1.4. describes the original performance of the outcome prediction scores described above. It includes the year of publication, the size of the development population, the moment in which the data are taken and their respective performance measures for hospital mortality discrimination. The predictive ability of the presented scores is measured with the Area Under the Receiver Operating Characteristic (AUROC) curve which is a performance measurement for binary classification problems that indicates how much model is capable of distinguishing between two classes. The AUROC returns a value between 0 and 1, the higher the AUROC the better the model. An AUROC close to 0 indicates that the model swapping the predictions (prediction ones as zeros and vice versa); an AUROC of 0.5 means that the model has no class separation capacity and an AUROC close to 1 indicates that the model has good measure of separability.

The AUROC performance presented in Table 1.4. shows the most recent versions of the APACHE score have a lower performance than previous version, however the test dataset is much bigger. In the case of SAPS scores, it can be observed that the SAPS III performance is lower that the SAPS II and the testing set it smaller, however, the important advance in this version is the reduction of data collection time (around two hours instead of the usual 24) and the inclusion of mortality prediction models for specific geographic regions. Despite the fact that OASIS is the most recent scoring system, among those reported, it is observed that its performance is the least, however, it is important to note that the objective of OASIS (13 variables) development was to create a score with discrimination and calibration similar of the APACHE IV (142 variables) but with less variables.

Table 1.4. Original performance of the outcome prediction scores.

Score	Score Version	Year published	Training dataset	Testing dataset	Selection of variables and their weights	Collection of data	Area Under Receiver Operating Characteristic curve (Hospital Mortality)
APACHE	APACHE	1981		805	Panel of experts	First 32 h in ICU	
	APACHE II	1985		5,815	Panel of experts	First 24 h in ICU	0.86
	APACHE III	1993	8,720	8,720	Multiple logistic regression	First 24 h in ICU	0.90
	APACHE IV	2006	66,335	44,223	Multiple logistic regression	First 24 h in ICU	0.88
SAPS	SAPS	1984		679	Panel of experts	First 24 h in ICU	
	SAPS II	1993	8,369	4,628	Multiple logistic regression	First 24 h in ICU	0.86
	SAPS III	2005	13,427	3,357	Multiple logistic regression	ICU admission ± 1h	0.85
OASIS	OASIS	2013	56,700	24,300	Genetic algorithm and Particle Swarm Optimization	First 24 h in ICU	0.837

Table 1.5. Studies which evaluated APACHE scores in an independent sample.

Score	Author	Country	Years	Patients	AUROC
APACHE II (22 studies)	Khwannimit	Thailand	2004-2005	1,316	0.888
	Vassar	US	1990-1991	2,414	0.87
	Ho	Australia	2005	1,311	0.858
	Ho	Australia	1993-2003	11,107	0.846
	Schneider	Australia and New Zealand	2001-2010	636,428	0.842
	Brinkman	Denmark	2006-2010	44,112	0.84
	Katsaragakis	Greece	1992-1997	661	0.839
	Beck	UK	1993-1996	16,646	0.835
	Markgraf	Germany	1991-1994	2,661	0.832
	Duke	Australia	2005-2007	1,843	0.82
	Nouira	Tunisia	1994-1995	1,325	0.82
	Peek	Netherlands	1999-2003	42,139	0.818
	Beck	UK	1993-1996	1144	0.806
	Capuzzo	Italy	1994-1997	1,721	0.805
	Harrison	UK	1995-2003	141,106	0.804
	Sakr	Germany	2004-2005	1,851	0.8
	Bastos	Brazil	1990-1991	1734	0.79
	Moreno	Portugal	1994-1995	982	0.787
	Livingston	Scotland	1995-1996	9,848	0.763
	Christensen	Denmark	2007	469	0.73
	Kim	Korea	2009	826	0.729
	Patel	US	1996-1997	302	0.702
APACHE III (14 studies)	Duke	Australia	2005-2007	1,843	0.91
	Zimmerman	US	1993-1996	37,668	0.89
	Vassar	US	1990-1991	2,414	0.89
	Paul	Australia and New Zealand	2004-2009	152,456	0.885
	Shann	Australia	2005-2006	16,356	0.88
	Keegan	US	2006	2,596	0.868
	Beck	UK	1993-1996	16,646	0.867
	Schneider	Australia and New Zealand	2001-2010	636,428	0.854
	Beck	UK	1993-1996	1144	0.847
	Markgraf	Germany	1991-1994	2,661	0.846
	Harrison	UK	1995-2004	141,107	0.832
	Pettila	Finland	1995	520	0.825
	Bastos	Brazil	1990-1991	1734	0.82
	Livingston	Scotland	1995-1996	10,326	0.795
APACHE IV (3 studies)	Kuzniewicz	US	1999-2003	11,300	0.892
	Brinkman	Denmark	2006-2009	55,661	0.87
	Keegan	US	2,006	2,596	0.861

Table 1.6. Studies which evaluated SAPS scores in an independent sample.

Score	Author	Country	Years	Patients	AUROC
SAPS II (30 studies)	Khwannimit	Thailand	2004-2005	1,316	0.911
	Soares	Brazil	2003-2005	952	0.88
	Kuzniewicz	USA	1999-2003	11,300	0.873
	Reiter	Austria	1998-2001	30,099	0.87
	Aegerter	France	1999-2000	13739	0.87
	Duke	Australia	2005-2007	1,843	0.87
	Katsaragakis	Greece	1992-1997	661	0.87
	LeGall	France	1998-1999	38,745	0.858
	Beck	UK	1993-1996	16,646	0.852
	Capuzzo	Italy	2006-2007	684	0.851
	Brinkman	Denmark	2006-2011	44,112	0.85
	Markgraf	Germany	1991-1994	2,661	0.846
	Nouira	Tunisia	1994-1995	1,325	0.84
	Peek	Netherlands	1999-2003	42,139	0.831
	Metnitz	Global	2002	16,784	0.83
	Haaland	Norway	2008-2010	10,135	0.83
	Poole	Italy	2007	3,661	0.83
	Metnitz	Austria	1997-1998	2,901	0.83
	Sakr	Germany	2004-2005	1,851	0.83
	Harrison	UK	1995-2005	141,108	0.822
	Moreno	Europe	1994-1995	10,027	0.822
	Strand	Norway	2006-2007	1,873	0.82
	Moreno	Portugal	1994-1995	982	0.817
	Capuzzo	Italy	1994-1997	1,721	0.816
	Metnitz	Austria	1997	1,733	0.81
	Apolone	Italy	1994	1393	0.8
	Livingston	Scotland	1995-1996	10,334	0.784
	Lim	Korea	2008-2009	633	0.76
	Christensen	Denmark	2007	469	0.74
	Patel	US	1996-1999	304	0.672
SAPS III (13 studies)	Khwannimit	Thailand	2007-2009	1,873	0.933
	Duke	Australia	2005-2007	1,843	0.88
	Soares	Brazil	2003-2005	952	0.87
	Silva Junior	Brazil	2008-2009	1,310	0.86
	Poole	Italy	2007	28,357	0.855
	Sakr	Germany	2004-2005	1,851	0.84
	Capuzzo	Italy	2006-2007	684	0.835
	Poole	Italy	2007	3,661	0.83
	Metnitz	Austria	2006-2007	2,060	0.82
	Strand	Norway	2006-2007	1,873	0.81
	Keegan	US	2006	2,596	0.801
	Lim	Korea	2008-2009	633	0.78
	Christensen	Denmark	2007	469	0.69

Table 1.7. Studies which evaluated OASIS scores in an independent sample.

Score	Author	Country	Years	Patients	AUROC
OASIS (3 studies)	Johnson	US	2001-2008	21,416	0.790
	Johnson	England	2007-2011	3,366	0.776
	Zhang	China	2012-2014	470	0.760

Several studies have analyzed the outcome prediction scores. In a recent PhD thesis by Johnson some performance studies are presented [13]. The inclusion criteria for a validation study was a cohort with at least 100 patients, a general ICU population (no disease specific evaluation) and evaluation of the model on an independent cohort. Table 1.5 and Table 1.6 summarizes the studies for APACHE and SAPS scores.

In addition, the external performance studies conducted in order to validate OASIS are presented in Table 1.7; OASIS uses a large multi-center population of ICU patients admitted to hospitals in the United States and additionally includes two external validation studies; the first one uses the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II database, a publicly available ICU database sourced from the Beth Israel Deaconess medical center in Boston, Massachusetts; and the second one takes place at a large tertiary teaching hospital in Oxford, England; an additional study that evaluates OASIS over an independent general ICU population of 470 patients is also presented [14].

The distribution of the AUROC performance for the studies is presented in Figure 1.1. It can be observed that there is no significate difference between the APACHE and SAPS scores. Figure 1.2 presents the number of different countries in which studies are reported, it can be observed that beside being the score that are appear in more performance studies, SAPS II is also, the score reported in a greater number of countries.

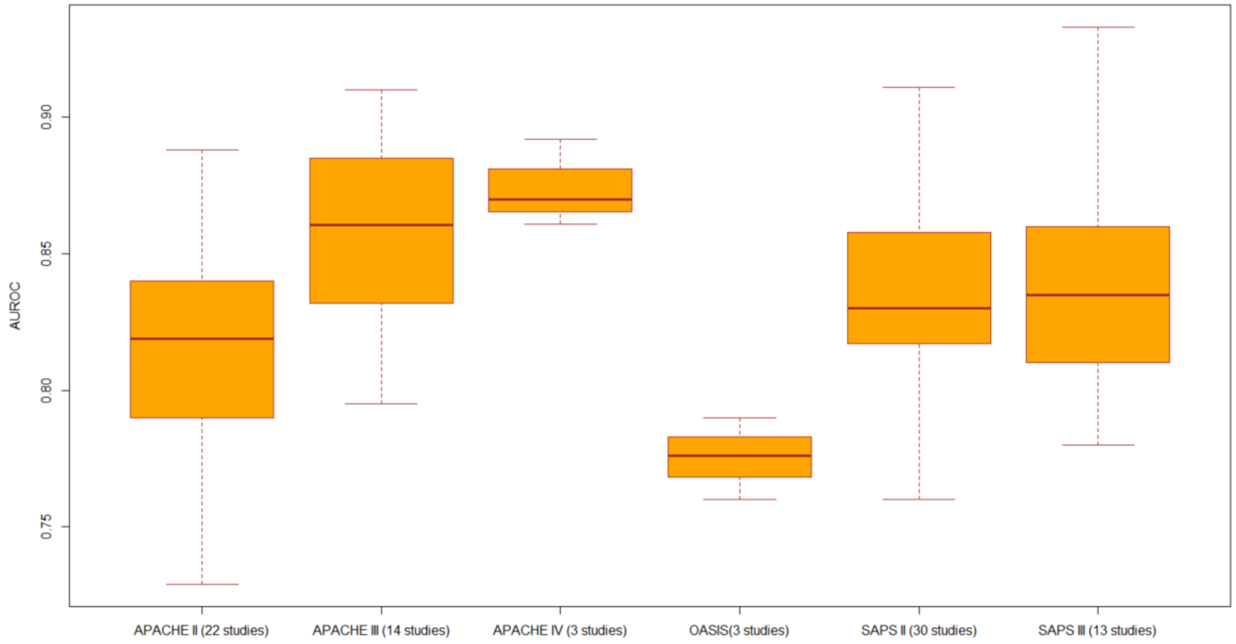


Figure 1.1. Boxplot of the AUROC performance of the studies on independent cohorts for APACHE, SAPS and OASIS scores.

1.2.1 Organ dysfunction scores

Other type of severity-of-illness scoring systems are designed to access the degree of organ dysfunction rather than to predict the outcome of a patient. This section describes two of the most several organ dysfunction scores used in ICU patients [7].

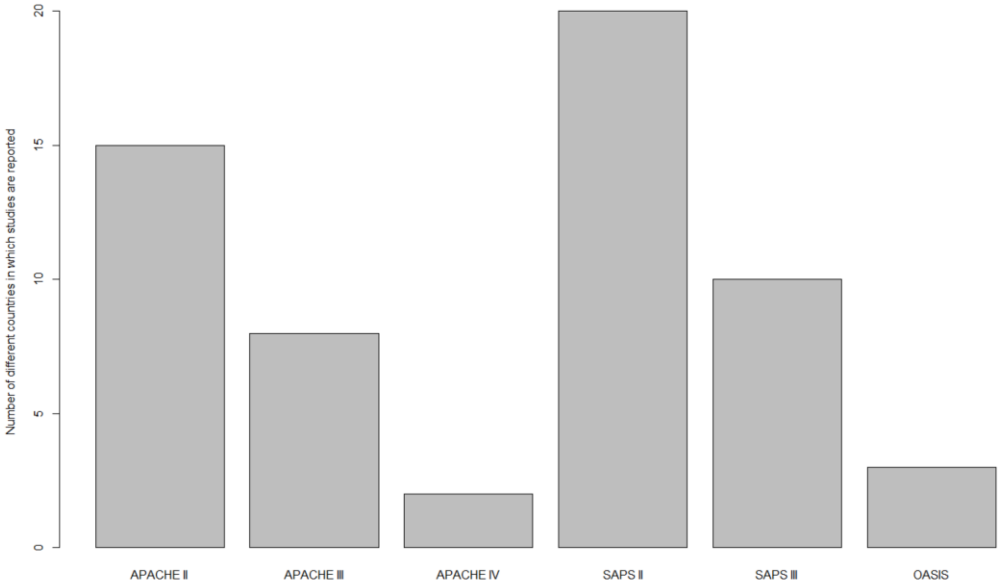


Figure 1.2. Number of different countries in which studies are reported.

1.2.1.1 Logistic Organ Dysfunction Score

The Logistic Organ Dysfunction Score (LODS) was developed in 1996, and uses a database of 13,152 admissions from 137 ICUs in 12 countries. LODS includes 12 variables to represent the function of six organ systems (neurologic, cardiovascular, renal, pulmonary, hematologic, hepatic). It takes the worst value for each variable in the first 24 hours of admission. Each system uses a score from 0 (that means no dysfunction) to 5 (that represents maximum dysfunction). Since LODS is a weighted system, it is possible to combine the total degree of organ dysfunction across the six organ systems in a single score that can be used as covariate in a logistic regression model to convert the global score into a probability of mortality [7, 8, 15]. Table 1.8 presents the LODS scoring system.

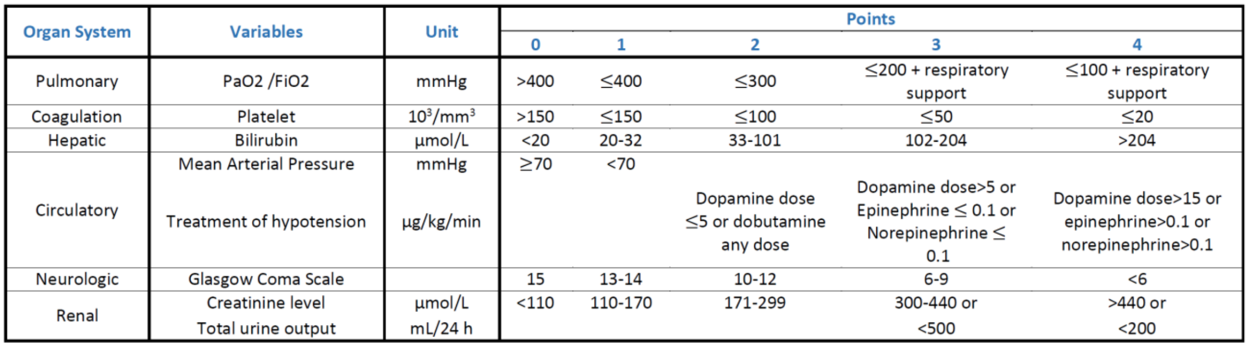
Table 1.8. Logistic Organ Dysfunction Score

Organ System	Variables	Unit	Points						
			5	3	1	0	1	3	5
Pulmonary	PaO2 /FiO2 in Mechanical ventilation or Continued positive airways pressure (CPAP)	%		< 150	≥150	With no ventilation, CPAP			
Hematologic	Platelet	$10^3/mm^3$			<50	≥50	≥50		
	Total leucocyte count	$10^3/mm^3$		<1.0	1.0-2.4	2.5-49.9			
Hepatic	Bilirubin	µmol/L				<34.2	≥34.2		
	Prothrombin time.	Seconds (s) and %			<25%	<3s or >25%	≥3s		
Cardiovascular	Systolic blood pressure;	mmHg	≥70	<70		≤239	240-269	≥270	
	Heart rate	bpm	<30	40-69	70-89	30-139			
Neurologic	Glasgow Coma Scale		3-5	6-8	9-13	14-15			
Renal	Creatinine level	µmol/L				<106.08	106.08-140.55	≥141.44	
	Total urine output	mL/24 h	<0.5	0.5-0.74		0.75-0.99			
	Ureic nitrogen	mmol/L				<6	6-9.98	9.99-19.98	≥19.99

1.2.1.2 Sequential Organ Failure Assessment

The Sepsis-related Organ Failure Assessment score was first developed in 1994; however, it eventually became known as the Sequential Organ Failure Assessment (SOFA) as it was applied outside of septic populations. Each function of the six organ systems (respiratory, cardiovascular, renal, hepatic, central nervous, coagulation) is scored from 0 (normal function) to 4 (most abnormal), which result in a possible score of 0 to 24. SOFA score takes the worst value on each day is recorded, and the cardiovascular component is assessed by a treatment-related variable (dose of vasopressor agents) instead of the composite variable [7, 8, 16]. Table 1.9 presents the SOFA scoring system.

Table 1.9. Sequential Organ Failure Assessment

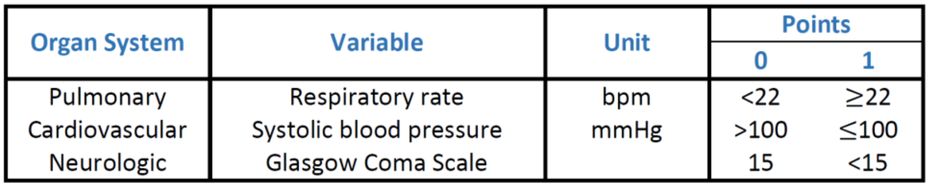


Organ System	Variables	Unit	Points				
			0	1	2	3	4
Pulmonary	PaO2 /FiO2	mmHg	>400	≤400	≤300	≤200 + respiratory support	≤100 + respiratory support
Coagulation	Platelet	10^3/mm^3	>150	≤150	≤100	≤50	≤20
Hepatic	Bilirubin	µmol/L	<20	20-32	33-101	102-204	>204
Circulatory	Mean Arterial Pressure	mmHg	≥70	<70			
	Treatment of hypotension	µg/kg/min			Dopamine dose ≤5 or dobutamine any dose	Dopamine dose>5 or Epinephrine ≤ 0.1 or Norepinephrine ≤ 0.1	Dopamine dose>15 or epinephrine>0.1 or norepinephrine>0.1
Neurologic	Glasgow Coma Scale		15	13-14	10-12	6-9	<6
Renal	Creatinine level	µmol/L	<110	110-170	171-299	300-440 or	>440 or
	Total urine output	mL/24 h				<500	<200

1.2.1.3 Quick Sequential Organ Failure Assessment

To facilitate recognition in prehospital, ward, and the emergency department, the Third International Sepsis Consensus Definitions Task Force [17] recommended a new severity of illness classification system, called "qSOFA" for quick sepsis-related organ dysfunction assessment score. The score ranges from 0 to 3 points. The presence of 2 or more qSOFA points near the onset of infection was associated with a greater risk of death or prolonged intensive care unit stay. Table 1.10 presents the qSOFA system.

Table 1.10. quick Sequential Organ Failure Assessment



Organ System	Variable	Unit	Points	
			0	1
Pulmonary	Respiratory rate	bpm	<22	≥22
Cardiovascular	Systolic blood pressure	mmHg	>100	≤100
Neurologic	Glasgow Coma Scale		15	<15

1.2.1.4 Organ dysfunction scores and new definition of sepsis

Recently, the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) update the definitions for sepsis and septic shock. The Task Force recommendations is that Sepsis should be defined as life-threatening organ dysfunction caused by a dysregulated host response to infection [17].

In the absence of a gold-standard diagnostic test for sepsis, the Task Force evaluates four clinical criteria in order to operationalize the new definition. They include SOFA, LODS, systemic inflammatory response syndrome (SIRS). the most important findings were: among ICU encounters with suspected infection, the predictive validity for in-hospital mortality of SOFA was not significantly different than the more complex LODS; but was statistically greater than SIRS and qSOFA. For this reason, the Sepsis-3 consensus recommend that for clinical operationalization, organ dysfunction can be represented by an increase in the SOFA score of 2 points or more which is associated with an in-hospital mortality greater than 10%. Among encounters with suspected infection outside of the ICU, the predictive validity for in-hospital mortality of qSOFA was statistically greater than SOFA and SIRS, supporting its use as a prompt to consider possible sepsis [18].

1.3 Mortality prediction in the ICU

Patient mortality is one of the most important clinical outcomes within an ICU. Traditionally, models associated with severity scores are used to assess the in-hospital mortality, such models are based on the analysis of large populations, and often provide statistically rigorous results for an average patient but are also expensive, time-consuming, and prone to selection bias, moreover, these indicators lack the precision required for use at the individual level and they yielded widely dissimilar performances when applied to different groups of patients, Table 1.11 presents as example six studies which compare severity score based models on different populations [19–24].

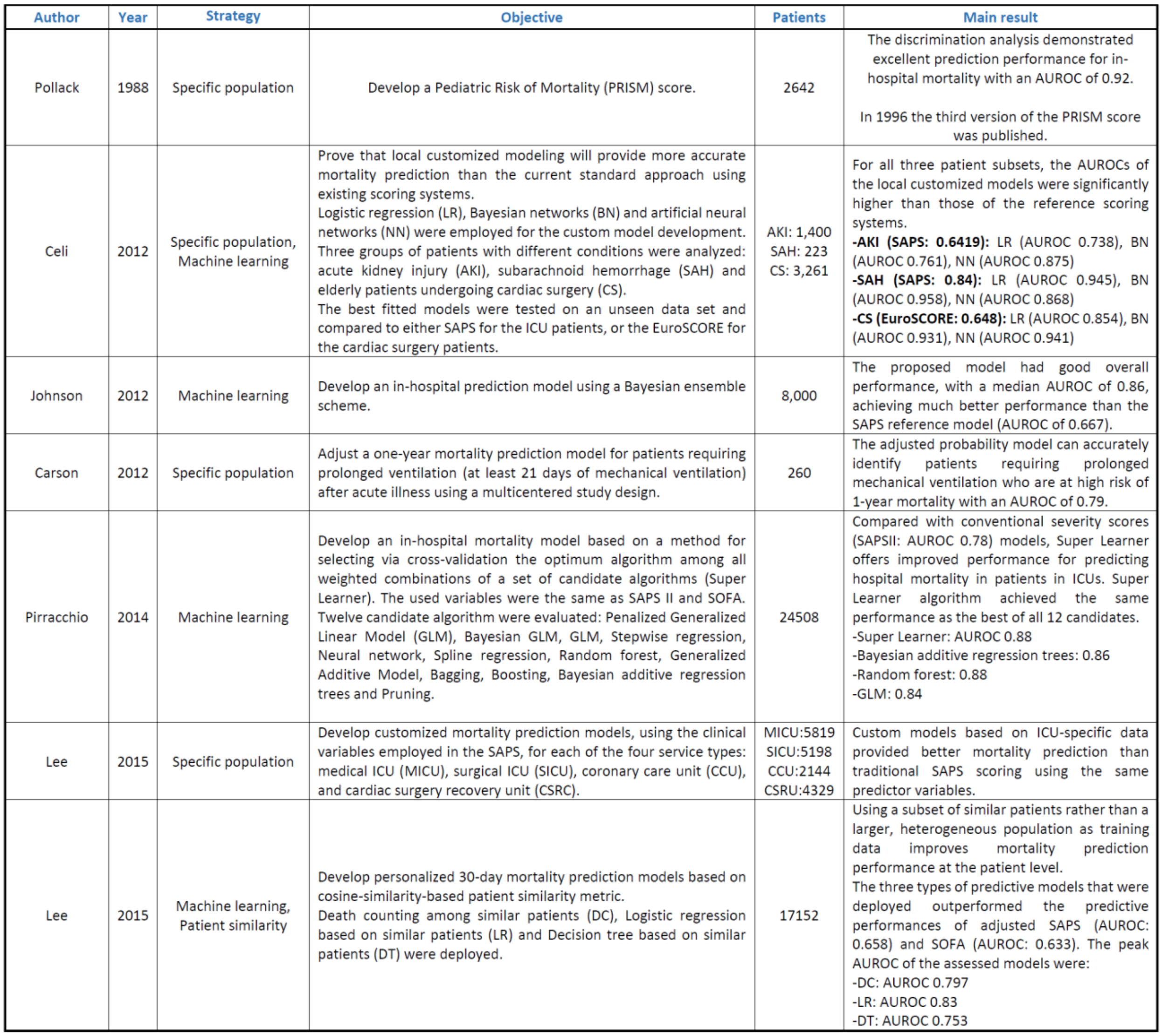
Table 1.11. Studies which evaluated Scoring Systems for disease specific samples.

Author	Year	Objective	Patients	Main result
Arabi	2003	Assess the validity of mortality prediction scoring systems in patients admitted to the ICU with severe sepsis and septic shock. The evaluated scores were: APACHE II, SAPS II, MPM II0 and MPM II24 and customized versions of SAPS II, MPM II24.	250	The overall mortality prediction was adequate for all six systems, discrimination was best for customized MPM II24 (AUROC: 0.826); the other performances are: MPM II24 (AUROC:0.823); MPM II0 (AUROC:0.806); customized SAPS II (AUROC:0.799); SAPS II (AUROC:0.797); and APACHE II (AUROC:0.782).
Hu	2013	Compare the efficiency of APACHE IV with that of MELD scoring system for prediction of the risk of hospital mortality after orthotropic liver transplantation	195	APACHE IV (AUROC: 0.937) showed better discrimination than MELD (AUROC: 0.694)
Haq	2014	Evaluate the utility of three ICU scoring systems: SAPS II, SAPS III, and APACHE II, to predict hospital mortality in patients above 90 years.	951	The SAPS III (AUROC: 0.81) presents the better performance followed by SAPS II(AUROC: 0.75) and APACHE II (AUROC: 0.74).
Saleh	2015	Compare the predictive accuracy of four predictive scoring systems in a single-center ICU subpopulation with acute respiratory distress syndrome: APACHE II, APACHE III, SAPS II and SOFA.	110	The accuracy of the studied scoring systems for predicting ICU mortality in acute respiratory distress syndrome patients is limited. The performance of the APACHE II/III scoring systems was superior to that of other systems in terms of predicting the mortality, and the combination of scores improved the performance.
Sun	2017	Determine the value of the SOFA, APACHE IV and SAPS II scoring systems in predicting short-term mortality of adult patients with acute myocarditis.	305	SAPS II (AUROC: 0.942) had the highest accuracy followed by the APACHE IV (AUROC: 0.934) and SOFA(AUROC: 0.920) scores.
Jentzer	2018	Determine the predictive value of outcome severity scores for mortality in cardiac intensive care unit patients who are 70 or older.	10,004 Patients <70 4,771 Patients ≥ 70	Severity of illness scores had lower AUROC values for hospital mortality in patients ≥ 70 years of age compared with patients <70 years of age. AUROC values were equivalent for the APACHE III (<70: 0.85; ≥70: 0.78), APACHE IV (<70: 0.86; ≥70: 0.79) and SOFA (<70: 0.86; ≥70: 0.80), and AUROC values for these scores were higher compared with OASIS (<70: 0.83; ≥70: 0.76).

It is important to highlight the work of Jentzer et al, since they present the largest published study comparing outcomes between patients ≥70 years of age with patients <70 years of age in a contemporary population. In such work, the authors concluded that AUROC values were equivalent for the APACHE-III, APACHE-IV and SOFA scores for in-hospital mortality, but it can also be observed that severity scores have lower discrimination for mortality in patients ≥70 than in older patients, which indicates that mortality prediction models based on traditional severity scores present errors at patient data away from the average.

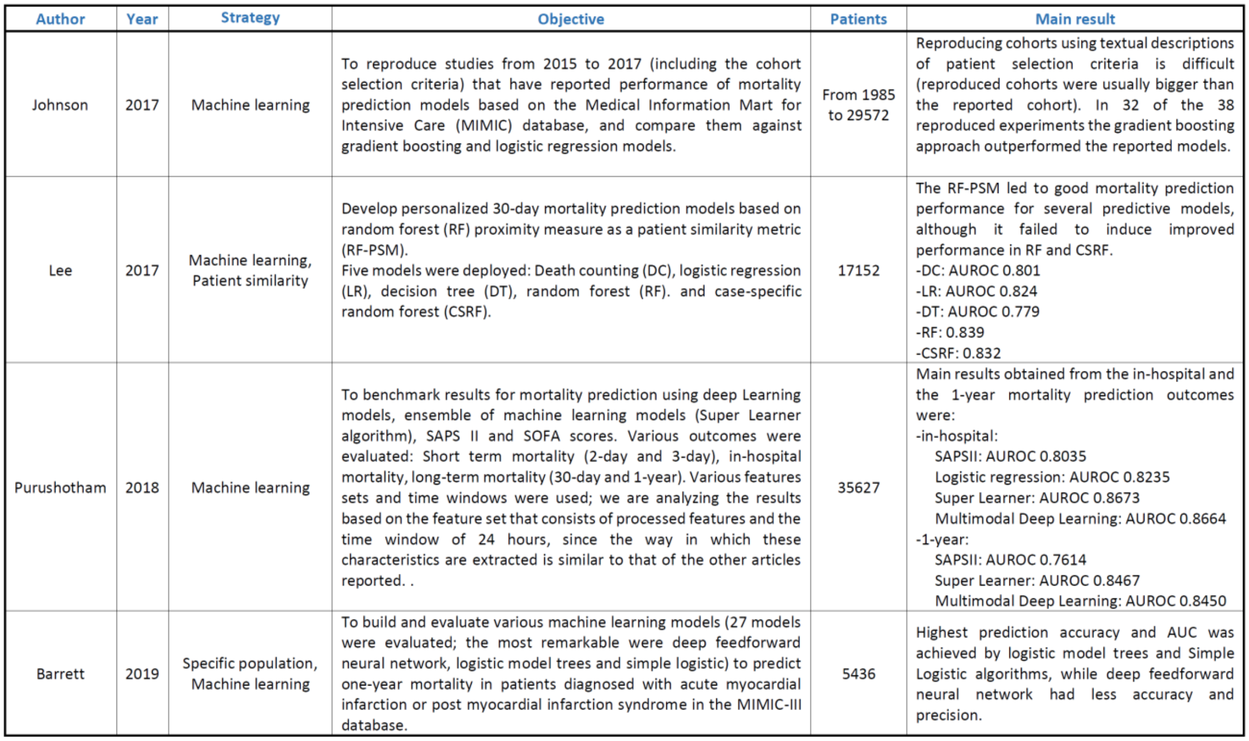
Efforts have been made to develop mortality prediction models with improved performance, and three strategies (That can be mixed with each other) stand out. The first one is based on the creation of specific models for groups of patients that shares a common characteristic (like diagnostics, service type or a similarity metric). The second one hinge on advanced machine learning algorithms different than logistic regression, and the third one relies on patient similarity. Table 1.12 presents some of these studies [1, 2, 25–32].

Table 1.12. Studies in the field of mortality prediction.



Author	Year	Strategy	Objective	Patients	Main result
Pollack	1988	Specific population	Develop a Pediatric Risk of Mortality (PRISM) score.	2642	The discrimination analysis demonstrated excellent prediction performance for in-hospital mortality with an AUROC of 0.92. In 1996 the third version of the PRISM score was published.
Celi	2012	Specific population, Machine learning	Prove that local customized modeling will provide more accurate mortality prediction than the current standard approach using existing scoring systems. Logistic regression (LR), Bayesian networks (BN) and artificial neural networks (NN) were employed for the custom model development. Three groups of patients with different conditions were analyzed: acute kidney injury (AKI), subarachnoid hemorrhage (SAH) and elderly patients undergoing cardiac surgery (CS). The best fitted models were tested on an unseen data set and compared to either SAPS for the ICU patients, or the EuroSCORE for the cardiac surgery patients.	AKI: 1,400 SAH: 223 CS: 3,261	For all three patient subsets, the AUROCs of the local customized models were significantly higher than those of the reference scoring systems. **-AKI (SAPS: 0.6419):** LR (AUROC 0.738), BN (AUROC 0.761), NN (AUROC 0.875) **-SAH (SAPS: 0.84):** LR (AUROC 0.945), BN (AUROC 0.958), NN (AUROC 0.868) **-CS (EuroSCORE: 0.648):** LR (AUROC 0.854), BN (AUROC 0.931), NN (AUROC 0.941)
Johnson	2012	Machine learning	Develop an in-hospital prediction model using a Bayesian ensemble scheme.	8,000	The proposed model had good overall performance, with a median AUROC of 0.86, achieving much better performance than the SAPS reference model (AUROC of 0.667).
Carson	2012	Specific population	Adjust a one-year mortality prediction model for patients requiring prolonged ventilation (at least 21 days of mechanical ventilation) after acute illness using a multicentered study design.	260	The adjusted probability model can accurately identify patients requiring prolonged mechanical ventilation who are at high risk of 1-year mortality with an AUROC of 0.79.
Pirracchio	2014	Machine learning	Develop an in-hospital mortality model based on a method for selecting via cross-validation the optimum algorithm among all weighted combinations of a set of candidate algorithms (Super Learner). The used variables were the same as SAPS II and SOFA. Twelve candidate algorithm were evaluated: Penalized Generalized Linear Model (GLM), Bayesian GLM, GLM, Stepwise regression, Neural network, Spline regression, Random forest, Generalized Additive Model, Bagging, Boosting, Bayesian additive regression trees and Pruning.	24508	Compared with conventional severity scores (SAPSII: AUROC 0.78) models, Super Learner offers improved performance for predicting hospital mortality in patients in ICUs. Super Learner algorithm achieved the same performance as the best of all 12 candidates. -Super Learner: AUROC 0.88 -Bayesian additive regression trees: 0.86 -Random forest: 0.88 -GLM: 0.84
Lee	2015	Specific population	Develop customized mortality prediction models, using the clinical variables employed in the SAPS, for each of the four service types: medical ICU (MICU), surgical ICU (SICU), coronary care unit (CCU), and cardiac surgery recovery unit (CSRC).	MICU:5819 SICU:5198 CCU:2144 CSRU:4329	Custom models based on ICU-specific data provided better mortality prediction than traditional SAPS scoring using the same predictor variables.
Lee	2015	Machine learning, Patient similarity	Develop personalized 30-day mortality prediction models based on cosine-similarity-based patient similarity metric. Death counting among similar patients (DC), Logistic regression based on similar patients (LR) and Decision tree based on similar patients (DT) were deployed.	17152	Using a subset of similar patients rather than a larger, heterogeneous population as training data improves mortality prediction performance at the patient level. The three types of predictive models that were deployed outperformed the predictive performances of adjusted SAPS (AUROC: 0.658) and SOFA (AUROC: 0.633). The peak AUROC of the assessed models were: -DC: AUROC 0.797 -LR: AUROC 0.83 -DT: AUROC 0.753

Table 12. (Continued). Studies in the field of mortality prediction.



Author	Year	Strategy	Objective	Patients	Main result
Johnson	2017	Machine learning	To reproduce studies from 2015 to 2017 (including the cohort selection criteria) that have reported performance of mortality prediction models based on the Medical Information Mart for Intensive Care (MIMIC) database, and compare them against gradient boosting and logistic regression models.	From 1985 to 29572	Reproducing cohorts using textual descriptions of patient selection criteria is difficult (reproduced cohorts were usually bigger than the reported cohort). In 32 of the 38 reproduced experiments the gradient boosting approach outperformed the reported models.
Lee	2017	Machine learning, Patient similarity	Develop personalized 30-day mortality prediction models based on random forest (RF) proximity measure as a patient similarity metric (RF-PSM). Five models were deployed: Death counting (DC), logistic regression (LR), decision tree (DT), random forest (RF). and case-specific random forest (CSRF).	17152	The RF-PSM led to good mortality prediction performance for several predictive models, although it failed to induce improved performance in RF and CSRF. -DC: AUROC 0.801 -LR: AUROC 0.824 -DT: AUROC 0.779 -RF: 0.839 -CSRF: 0.832
Purushotham	2018	Machine learning	To benchmark results for mortality prediction using deep Learning models, ensemble of machine learning models (Super Learner algorithm), SAPS II and SOFA scores. Various outcomes were evaluated: Short term mortality (2-day and 3-day), in-hospital mortality, long-term mortality (30-day and 1-year). Various features sets and time windows were used; we are analyzing the results based on the feature set that consists of processed features and the time window of 24 hours, since the way in which these characteristics are extracted is similar to that of the other articles reported. .	35627	Main results obtained from the in-hospital and the 1-year mortality prediction outcomes were: -in-hospital: SAPSII: AUROC 0.8035 Logistic regression: AUROC 0.8235 Super Learner: AUROC 0.8673 Multimodal Deep Learning: AUROC 0.8664 -1-year: SAPSII: AUROC 0.7614 Super Learner: AUROC 0.8467 Multimodal Deep Learning: AUROC 0.8450
Barrett	2019	Specific population, Machine learning	To build and evaluate various machine learning models (27 models were evaluated; the most remarkable were deep feedforward neural network, logistic model trees and simple logistic) to predict one-year mortality in patients diagnosed with acute myocardial infarction or post myocardial infarction syndrome in the MIMIC-III database.	5436	Highest prediction accuracy and AUC was achieved by logistic model trees and Simple Logistic algorithms, while deep feedforward neural network had less accuracy and precision.

Three of the reported works, evaluated various machine learning models on the same cohort an assessed mortality prediction. Pirracchio et Al. [27] used a Super Learner (SL) algorithm to predict in-hospital mortality, SL is a supervised learning algorithm that is designed to find the optimal combination from a set of prediction algorithms, meaning, that the SL is an ensemble machine learning technique that uses multiple learning algorithms to obtain better prediction performance, and in theory it perform at least as well as the best member of the library of prediction algorithms that it uses. In their work, Pirracchio et Al. used twelve prediction algorithm: Penalized Generalized Linear Model (GLM), Bayesian GLM, GLM, Stepwise regression, Neural network, Spline regression, Random forest, Generalized Additive Model, Bagging, Boosting, Bayesian additive regression trees and Pruning, which means that, in order to build the SL, the performances of each of the twelve prediction algorithms were assessed, and, in a test done with non-processed variables, it was found that best performing algorithm was the random forest and the worst performing one was neural network.

Purushotham et Al. [30] benchmarked the performance of the deep learning models with respect to traditional severity scoring systems and the super learner described by Pirracchio et Al. on the Medical Information Mart for Intensive Care III (MIMIC-III) database. The evaluated deep learning models were: Feedforward neural networks (FFN), recurrent neural network (RNN) and multimodal deep learning model (MMDL). In their work Purushotham et Al. used various mortality prediction benchmark tasks: In-hospital mortality, short-term mortality (2-day and 3-day mortality) and long-term mortality (30-day and 1-year); they also selected three sets of features: Feature set A, consisting of the 17 processed features used in the calculation of the SAPS-II score; feature set B, consisting of 20 raw values related to the SAPS-II score and feature set C, consisting of 136 raw features. For all the mortality prediction tasks evaluated with feature set A there is not much difference between the MMDL model and the super learner performances, however when raw features are used the MMDL model consistently obtains the best results. Indicating that deep learning models benefit from large number of raw features.

Barret et Al. evaluated various machine learning models, including a deep learning model on a specific population: ICU patients with acute myocardial infarction and post myocardial infarction syndrome [31]. In their study Barret et Al. analyzed data from 5436 patients and found out that, from the 27 used algorithms, the best performing ones were logistic model trees and simple logistic model.

Another remarkable study was presented by Johnson et Al. [28]. and focus on the reproducibility of mortality prediction studies within the ICU. In their study Johnson et Al. reproduced 38 experiments that use MIMIC database, and compare the performance reported in the studies against gradient boosting and logistic regression models using a simple set of features. The outcome for prediction was define by the study, and was one of the following: in-hospital mortality, 30-day post ICU admission mortality, 48-hour post ICU discharge mortality, 30-day post ICU discharge mortality, 30-day post hospital discharge mortality, 6-month post hospital discharge mortality, 1-year post hospital discharge mortality, and 2-year post hospital discharge mortality. The cohorts were also defined by the study, since Johnson et Al. attempted to reproduce each of the cohorts used in the studies. Unlike the studies reported above, in which multiple machine learning algorithms were evaluated on the same cohort, in this work two algorithms were evaluated on multiple cohorts with the same features. On the 38 reproduced experiments all the non-linear prediction models (gradient boosting) outperformed the linear prediction models (logistic regression), and, in average, the discrimination performance difference was 2.42%, which indicates that the mortality prediction can be approached quite linearly. On the other hand, the gradient boosting exhibited better results than those published in the review studies in 32 of the 38 experiments; and the logistic regression showed better results in 27 experiments, 16 of which were approached with non-linear models in the original studies.

It can also be observed that in the field of personalized predictive modeling based on patient similarity for mortality prediction in the ICU, Lee et Al. deployed a cosine-similarity-based patient similarity metric to identify patients that are most similar to an index patient and subsequently custom-build a 30-day mortality prediction model which outperformed the results obtained with models fitted with all the data and traditional severity of disease scores [2], in their experiments 5000 was determined to be the minimum number of similar patients for logistic regression to ensure sufficient variability in categorical predictors within training data (these minimum numbers of similar patients could be different for other datasets and predictors) and the best performance (highest AUROC) were achieved with logistic regression when 5000 or 6000 most similar patients were used for training the personalized model. One of the main conclusions of this work is that using a subset of similar patients rather than a larger, heterogeneous population as training data improves mortality prediction performance at the patient level. In this study, predictors equally contribute to the patient similarity metric, the patient cohort is a representation of patients with a wide variety of diagnoses and conditions and a personalized model is fitted for each index admission. In a later work, Lee et Al. [29] used a random forest proximity measure as a patient similarity metric in the context of personalized mortality prediction within the ICU, this work used the same population and methodology that their previous one, and it can be observed that, in comparison with the death counting (DC), logistic regression (LR), and decision trees (DT) results from their previous work that studied a cosine similarity as patient similarity metric, the predictive performance was similar, moreover, random forest and case-specific random forests did not benefit from personalization via the use of the random forest patient similarity metric. The above, and the fact that the random forest modeling approaches did not benefit from personalization via the use of the random forest similarity measure, indicate that selecting an appropriate similarity metric is not a straightforward task.

1.4 Conclusion

General severity of illness scores can be useful to guide prognostication, to assess ongoing disease development and organ function, to compare ICU performance over time and across units, to compare clinical trial population and outcomes. The general traditional scores were developed to be used in mixed groups of ICU patients, for this reason their accuracy in subgroups of patients can be questioned; even more since the begging of severity of illness scores it is clear the need to create specific scoring systems according to the characteristics of the patients, that is why scores like APACHE and SAPS were developed for adult population while the PRISM was constituted as a pediatric score and the SAPS III include specific coefficients to geographic regions.

In a survey Bouch listed the characteristics for an ideal scoring system [33]:

1. On the basis of easily/routinely recordable variables
2. Well calibrated
3. A high level of discrimination
4. Applicable to all patient populations
5. Can be used in different countries
6. The ability to predict functional status or quality of life after ICU discharge.

Studies presented in Table 1.11 suggest that no scoring system currently incorporates all these features, specifically the items 4 and 5 are challenging to fulfill, that is why machine learning approaches and disease-specific scoring systems, like the ones presented in the Table 1.12, are increasingly being developed. Specifically, from the works that evaluated multiple models it can be interpreted that deep learning models require large training and feature sets to report improvements, which can be seen in the fact that in the Barret et Al. study [31] (performed on a specific population) the simple logistic and logistic trees methodologies outperformed the deep learning models; but in the Purushotham et Al. [30] study (performed on a general population) the deep learning models exhibited the best results. In addition to this, it can also be observed from Table 1.12 that the ensemble methodologies based on trees (like random forest and gradient boosting) consistently report good performances.

The customize models have proved to perform better than the general population approach, however these studies continue to be population-based and therefore they generally provide “the average best choice”. One developing idea in this field is personalized predictive modeling based on patient similarity. The goal of this approach is to identify patients who are similar to an index patient and derive insights from the data of similar patients to provide personalized predictions. This approach has been widely used for personalized predictions in other fields, including music, movies and e-commerce, however, there are still very few studies that focus on personalized prediction driven by patient similarity metrics within the ICU.

CHAPTER 2. SEPSIS

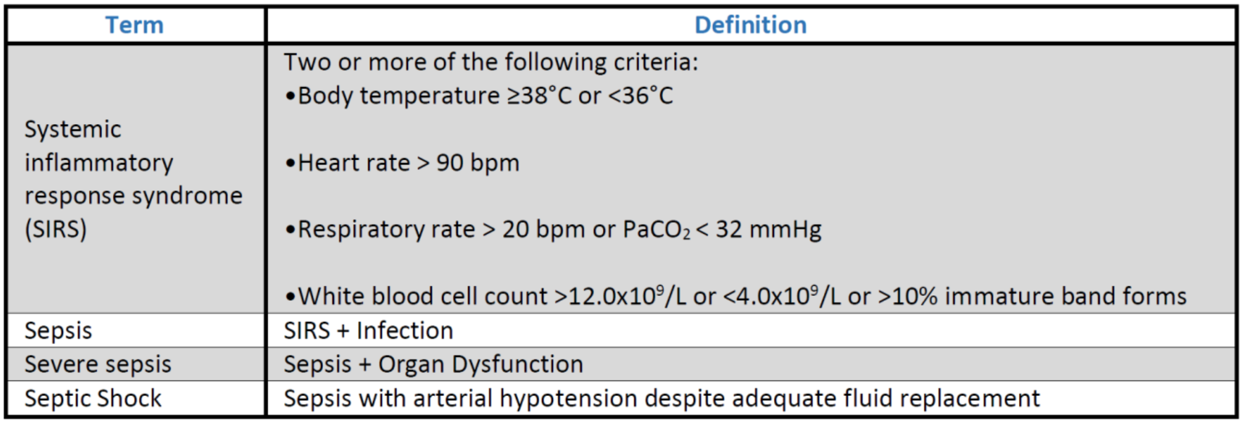
2.1 Introduction

In this chapter we seek to show how the outcome prediction problem for sepsis patients within the ICU is currently being addressed, for this, we first present the current definition of sepsis, the incidence of the condition and the estimates for its mortality. We also present four criteria for sepsis identification based on retrospective analysis of ICD codes and administrative data generated at the end of the hospital stay. Then we indicate the long-term outcomes for patients with sepsis, and finally we present studies focused on the mortality prediction exclusively for sepsis patients is within the ICU.

Sepsis is a word derived from the ancient Greek [σηψις], which means the decomposition of animal- or plant-based organic materials by bacteria. The modern concept was introduced in 1991 in a consensus conference held by the American College of Chest Physicians (ACCP) and the Society of Critical Care Medicine (SCCM) where sepsis was defined as the host's inflammatory response to infection [34, 35].

From that moment sepsis was considered a condition resulted from a host’s systemic inflammatory response syndrome (SIRS) to infection. When organ dysfunction occurred, it was considered severe sepsis, a condition that, if aggravated, could turn into septic shock, defined as “sepsis-induced hypotension persisting despite adequate fluid resuscitation.” [36, 37]. Table 2.1 presents the summary of definitions.

Table 2.1 Definitions for systemic inflammatory response syndrome (SIRS), sepsis, severe sepsis, and septic shock



Term	Definition
Systemic inflammatory response syndrome (SIRS)	Two or more of the following criteria: •Body temperature ≥38°C or <36°C •Heart rate > 90 bpm •Respiratory rate > 20 bpm or $PaCO_2 < 32$ mmHg •White blood cell count >12.0x10^9/L or <4.0x10^9/L or >10% immature band forms
Sepsis	SIRS + Infection
Severe sepsis	Sepsis + Organ Dysfunction
Septic Shock	Sepsis with arterial hypotension despite adequate fluid replacement

The SIRS criteria have been used for identification of potentially septic patients because they can facilitate enrollment for research purposes; however, their utility is limited by the lack of specificity since up to 90% of patients admitted to the ICU fit the criteria for SIRS [38]. The above, and the advances into the pathobiology, management, and epidemiology of sepsis led to the reexamination of the definitions. As consequence, in The Third International Consensus Definitions for Sepsis and Septic Shock, a task force proposed a new definition that incorporate the current understanding of sepsis biology, defining sepsis as a “life-threatening organ dysfunction caused by a dysregulated host response to infection”[17, 18].

Under this new definition, sepsis involves organ dysfunction, indicating a pathobiology more complex than infection with an accompanying inflammatory response. It makes the term “severe sepsis” superfluous

[17], and septic shock is defined as a subset of sepsis with profound circulatory, cellular, and metabolic abnormalities [39].

For clinical operationalization, sepsis can be diagnosed when organ dysfunction happens, represented by an increase in the Sequential Organ Failure Assessment (SOFA) score of 2 points or more, consequent to an infection [40]. Septic shock can be identified using the clinical criteria of hypotension requiring vasopressor therapy to maintain mean arterial pressure (MAP) at least of 65mmHg and having a serum lactate level greater than 2 mmol/L after adequate fluid resuscitation [39]. Table 2.2 summarizes these definitions.

Table 2.2 New definitions for sepsis, organ dysfunction, and septic shock.

Term	Definition
Sepsis	Life-threatening organ dysfunction caused by a dysregulated host response to infection
Organ dysfunction	Change in total SOFA score≥2 points consequent to the infection
Septic Shock	Subset of sepsis patients with persisting hypotension requiring vasopressors to maintain MAP≥65mmHg and having a serum lactate level >2mmol/L despite volume resuscitation.

2.2 Sepsis criteria

To operationalize the new definition of sepsis, presented in the previous section, the task force of Third International Consensus Definitions for Sepsis and Septic Shock recommends to replace the SIRS criteria with the SOFA score [40].However, despite the efforts of the task force for standardize a sepsis diagnostic, there remains some controversy around the new definitions [41–46]: new definitions did not involve low or middle income countries; as result, SOFA is a score is routinely calculated in some, but not all, ICUs; the decision of replace SIRS with SOFA was based on a retrospective study conducted with ICU patients with sepsis in which it was observed that 1 out of 8 patients with sepsis and multiorgan failure did not have at least 2 SIRS criteria, and, semantically, it cannot be ignored that in 7 out of 8 cases the patients met the “at least 2 SRIS” criteria. Even the experts in sepsis pathobiology of the consensus recognized some limitations since some of the definitions and clinical criteria were generated through voting, and unanimity was not always presented.

In a respond to the debate, Singer (one of the specialist of the sepsis-3 task force), argues that the main reason why SIRS was not included in the operationalization of the new sepsis definition was based on pathophysiology, because SIRS criteria are not particularly good in distinguishing a normal and appropriate host response to an infection from an inappropriate response resulting in a more serious infection [41].

Sepsis-3 definition appears to be an improvement over the previous iterations and the main purpose of using SOFA score for operationalizing sepsis is to diagnose the condition. In contrast to SIRS, the new definition is all-inclusive as it reflects new onset organ dysfunction. On the other hand, for retrospective studies it is possible to identify sepsis by using different criteria that uses ICD codes and administrative data generated at the end of the hospital stay.

2.2.1 Explicit sepsis

The current definition of sepsis indicates that the condition is life-threatening organ dysfunction caused by a dysregulated host response to infection, for this reason the codes of the International Classification of Diseases (ICD) which best frame the new definition are 995.92 for severe sepsis and 785.52 for septic shock. These codes are extremely specific to sepsis, but have very low sensitivity.

2.2.2 Angus criteria

The Angus criteria is a validated protocol that uses administrative data to identify sepsis patients. The algorithm for the Angus [47] criteria first looks to identify patients coded for severe sepsis or septic shock. If patients do not have this code, all discharge diagnoses are reviewed for an infection code, if present then procedure codes/diagnoses codes are checked for organ dysfunction codes [38]. Table 2.3 presents the ICD-9-CM-based classification of acute organ dysfunction used by the Angus Criteria.

Table 2.3. Angus ICD-9-CM-based classification of acute organ dysfunction. Where 3- or 4-digit codes are listed, all associated subcodes were included.

Organ System	ICD-9-CM Code	ICD-9-CM Code Description
Cardiovascular	785.5	Shock without trauma
	458	Hypotension
Respiratory	96.7	Mechanical ventilation
Neurologic	348.3	Encephalopathy
	293	Transient organic psychosis
	348.1	Anoxic brain damage
Hematologic	287.4	Secondary thrombocytopenia
	287.5	Thrombocytopenia unspecified
	286.9	Other/unspecified coagulation defect
	286.6	Defibrination syndrome
Hepatic	570	Acute and subacute necrosis of liver
	573.4	Hepatic infarction
Renal	584	Acute renal failure

2.2.3 Martin criteria

The criteria developed by Martin et al.[48] sorts patients either by codes for septicemia, septicemic, bacteremia, disseminated fungal infection, disseminated candida infection or disseminated fungal endocarditis in addition to an organ dysfunction code or an explicit diagnosis: severe sepsis or septic shock [38]. Table 2.4 presents the ICD-9-CM-based classification of acute organ dysfunction used by the Martin Criteria.

2.2.4 Sepsis-3

In a recent study Desautels et al. aimed to study and validate a sepsis prediction method, for the new Sepsis-3 definitions based on retrospective data [49]. For this they took the earliest culture draw or antibiotic administration as the time of suspicion of infection, and then they define a window of up to 48 hours before this time and 24 hours after this time. The SOFA score, at the beginning of this window, was compared with its hourly value throughout this window; and when the hourly value was ≥ 2 points higher than the value at the start of the window the particular admission was designate as septic.

Table 2.4. Martin ICD-9-CM Based Classification of Acute Organ Dysfunction Associated with Sepsis.

Organ System	ICD-9-CM Code	ICD-9-CM Code Description
Respiratory	518.81	Acute respiratory failure
	518.82	Acute respiratory distress syndrome
	518.85	Acute respiratory distress syndrome after shock or trauma
	786.09	Respiratory insufficiency
	799.1	Respiratory arrest
	96.7	Ventilator management
Cardiovascular	458	Hypotension, postural
	785.5	Shock
	785.51	Shock, cardiogenic
	785.59	Shock, circulatory or septic
	458	Hypotension, postural
	458.8	Hypotension, specified type, not elsewhere classified
	458.9	Hypotension, arterial, constitutional
	796.3	Hypotension, transient
Renal	584	Acute renal failure
	580	Acute glomerulonephritis
	585	Renal shutdown, unspecified
	39.95	Hemodialysis
Hepatic	570	Acute hepatic failure or necrosis
	572.2	Hepatic encephalopathy
	573.3	Hepatitis, septic or unspecified
Hematologic	286.2	Disseminated intravascular coagulation
	286.6	Purpura fulminans
	286.9	Coagulopathy
	287.3	Thrombocytopenia, primary, secondary, or unspecified
Metabolic	276.2	Acidosis, metabolic or lactic
Neurologic	293	Transient organic psychosis
	348.1	Anoxic brain injury
	348.3	Encephalopathy, acute
	780.01	Coma
	780.09	Altered consciousness, unspecified
	89.14	Electroencephalograph

In this work, we will use these four criteria to identify the admissions that will be part of our study cohort. Since each of these identification methodologies are based on different ICD-9 codes for the determination of organ dysfunction, it is expected that they yield widely different patient groups; however they all fulfill the current definition of sepsis: "life-threatening organ dysfunction caused by a dysregulated host response to infection", making them appropriate for our study.

2.3 Global Incidence and Mortality of Hospital-treated Sepsis

So far we have presented the definition of sepsis, and how the condition can be identified both at the time of attention and retrospectively, however, we have not yet seen why sepsis is considered a delicate health problem. In this chapter we present the incidence and mortality of hospital-treated sepsis elucidating why its study is relevant.

At the annual congress of the European Society of Intensive Care Medicine (ESICM) in October 2002, the Surviving Sepsis Campaign was formed with the objective to raise awareness to reduce sepsis mortality. At that time, sepsis was considered a leading cause of death in the intensive care unit with a worldwide documented incidence of 1.8 million each year; however, it was considered that this number is confounded by a low diagnostic rate and difficulties in tracking sepsis in many countries, and the Surviving Sepsis Campaign estimates that with an incidence of 3 in 1000 the true number of cases each year could reach 18 million, and the mortality rate was of almost 30% [50].

The Surviving Sepsis Campaign outlined a six-point action plan (Table 2.5) aimed at improving the management of sepsis, and increase awareness among health care professionals, governments, the public and funding bodies.

Table 2.5. The Surviving Sepsis six-point action plan

Issue	Details
Awareness	Increase awareness of health care professionals, governments, health and funding agencies, and the public of the high frequency and mortality associated with sepsis
Diagnosis	Improve the early and accurate diagnosis of sepsis by developing a clear and clinically relevant definition of sepsis and disseminating it to our peers
Treatment	Increase the use of appropriate treatments and interventions by disseminating the range of care options and urging their timely use
Education	Encourage the education of all health care professionals who manage sepsis patients by providing leadership, support and information to them about all aspects of sepsis management, including diagnosis, treatments and interventions, and standards of care
Counselling	Provide a framework for improving and accelerating access to post-ICU care and counselling for sepsis patients
Referral	Recognize the need for clear referral guidelines that are accepted and adopted at a local level in all countries by initiating the development of global guidelines

Existing epidemiologic studies suggest that sepsis remains a huge burden across all regions, despite that a sustained Surviving Sepsis Campaign achieved a continuous quality improvement in sepsis care [51]. Sepsis incidence rates are up to 535 cases per 100,000 person-years and rising [52], and although outcomes have improved, in-hospital remains high at 25–30% [52]. Patients with sepsis are frequently cared for in ICUs, either because sepsis itself led to their admission or because sepsis developed as a complication of their admission.

A recent metaanalysis by Fleischmann et al. estimated the worldwide incidence and mortality of sepsis; they systematically searched 15 international citation databases for population-level estimates of sepsis incidence rates and mortality in adult populations published in the last 36 years [53]. One of the main findings reported is that studies on population level incidence and mortality rates for sepsis and severe sepsis are scarce, and none exist for low- and middle-income countries. For the High-income countries (when the only the last 13 years were analyzed) an aggregate global estimator of 437 sepsis cases per 100,000 person-years is reported. For the case of severe sepsis population incidence rate of 270 cases per 100,000 person-years was estimated when more recent investigations were analyzed. The mortality rates estimated from hospital-treated cases are 17% for sepsis and 26% for severe sepsis from studies published between 2003 and 2015.

This estimated indexes only covers the high-income countries, which only represent 13% of the world's population. If the reported rates also apply for the countries low- and middle-income countries, a total annual number of 31.5 million sepsis and 19.4 million severe sepsis cases would be expected to be treated in hospitals around the globe each year, and it may cause or contribute to up to 5.3 million deaths worldwide per annum. However, the true incidence and burden of sepsis in low- and middle-income countries remains uncertain because of a lack of information on sepsis epidemiology and may even be higher since infectious diseases are considerably more prevalent in these areas of the world and cause a substantially higher proportion of deaths [53].

Specifically for Colombia, in 2011, Rodríguez et al. published a study with the aim of determine the epidemiologic characteristics of sepsis in a hospital based population in Colombia. This study was carried out using data from ten general hospitals in the four main cities of Colombia between September 1, 2007 and February 29, 2008. A total of 2,681 patients were recruited from emergency rooms, intensive care units, and general wards. The 28-day mortality rates of patients with infection without sepsis, sepsis without organ dysfunction, severe sepsis without shock, and septic shock were 3%, 7.3%, 21.9%, and 45.6%, respectively; study reports a monthly incidence of 3.61 cases of sepsis per 100 hospital admissions [54].

In a secondary analysis on the same cohort, Ortíz et al. focused on the patients admitted to the intensive care units and reported an overall 28-days mortality rate at the time of discharge of 33.6%, and could go up to 45.1% for patients with septic shock [55].

2.4 Long-term Outcomes from Sepsis

The data presented in the previous section focus on the short-term mortality; however, in recent years interest in understanding the impact of critical illness on long-term outcomes has increased. Studies examining long-term outcomes of severe sepsis, acute lung injury, and lung transplantation suggest that critical illness is associated with long-term consequences that persist beyond ICU and hospital stay [56].

In 2003 Weycker et al. [57] estimated the long-term mortality and medical care charges among patients with severe sepsis and concluded that their mortality and economic costs are high, during the period of acute illness as well as subsequently. The study identified 16,019 patients who were treated in hospital for severe sepsis and reported that 21.2% of subjects died in hospital, 51.4% died after one year and 74.2%

died after five years; with respect to medical charges, the mean total charges for the index admission were $44,600 (USD), at 1 year, mean cumulative medical care charges totaled $78,500 (USD); at 5 years, the total was $118,800 (USD).

In [56], authors suggested that mechanisms underlying increased long-term mortality remain poorly understood, and besides the fact that long-term mortality following severe sepsis is high, and fewer than half of patients who experience severe sepsis are alive at 1 year, other outcomes like neurologic impairments, respiratory impairment and renal failure are also important because they may increase risk of death and reduce quality of life.

In 2010, Iwashyna et al. [58], reported that severe sepsis was associated with substantial and persistent cognitive impairment and functional disability among survivors. And Winters et al [59] concluded that patients with sepsis showed ongoing mortality up to 2 years and beyond after the standard in hospital mortality end point. Patients with sepsis also had decrements in quality-of-life measures after hospital discharge.

In 2013 Wang et al. [60] concluded that sepsis is independently associated with increased risk of mortality after hospital treatment; Individuals with the disease exhibited increased rates of death for up to 5 years after the illness event. This is evidenced by the fact that One-year, 2-year and 5-year mortality among individuals with sepsis were 23%, 28.8% and 43.8%, respectively; and the death rates in the same periods of those patients who never developed sepsis were 1%, 2.6% and 8.3%.

In 2015 Ou et al [61] reported that sepsis survivors had substantially increased risks of all-cause mortality, as well as major adverse cardiovascular events like ischemic stroke, hemorrhagic stroke, myocardial infarction, heart failure, and sudden cardiac death or ventricular arrhythmia at 1 year after discharge when compared with matched population control subjects; sepsis survivors had higher risks.

2.5 Outcome prediction for sepsis patients in the intensive care unit

From chapter 1, it is clear that, traditionally approaches to ICU outcome prognostication has relied on static models generated from analyzing large, heterogeneous, multi-center patient datasets, such one-size-fits-all approaches perform well for the average patient; but tend to present problems when the characteristics of the patients move away from the mean. According to this, in section 1.3, we showed that efforts have been made to generate mortality prediction models that use data from patients who share the same characteristic (for example, the same diagnosis).

The mortality prediction problem exclusively for sepsis patient within the ICU has been addressed mainly with two approaches: One approach is to modify the models by adapting them for use specifically among patients with sepsis. The second one is to develop entirely new models, incorporating additional variables that could potentially enhance accuracy.

The first approach was proposed by Le Gall et al. in 1995 and then assessed by Arabi et al. in 2003 [62, 63]. In his 1995 work, Le Gall showed that SAPS II and Mortality Prediction Model II at 24 hours (MPM$_{24}$ II) did not fit the data well when used exclusively on severe sepsis patients, and proposed a methodology for the adjustment of those models to the severe sepsis and septic shock population. To adjust of these two scores for patients with severe sepsis, Le Gall et al. developed new logistic regression equations using only either the SAPS II or the MPM II$_{24}$ scores obtained from the group of patients with severe sepsis. Consequently, each new logistic regression model would contain a single variable plus the constant term.

The idea behind this approach is that the original score, which produced a probability of hospital mortality for general ICU population, would be mathematically translated into an adjusted probability of mortality based only on the experience of patients with severe sepsis. This adjusted versions of the SAPS II and MPM II_{24} presented better discrimination and calibration than the original models.

The performance of the adjusted mortality prediction scoring system proposed by Le Gall. Et al, and another four scores (APACHE II, SAPS II, Mortality Prediction Model II at admission (MPM II_0), MPM II_{24}) were evaluated in a cohort of 250 patients with suspected severe sepsis and septic shock by Arabi et al. [62]. They concluded that the overall mortality prediction was adequate for all six scores, however, calibration was inadequate for APACHE II, SAPS II, MPM II_0 and MPM II_{24}. On the other hand, the adjusted version of SAPS II and MPM II_{24} exhibited improved calibration that the original versions.

In addition to the adjustment of existing models, particular scores for the prediction of mortality in patients with severe sepsis and septic shock have been developed [64, 65]. Carrara et al. presented a development of a model exclusively for septic shock patients derived from hemodynamic variables, clinical information and laboratory results of the first 48 hours after shock onset and to predict mortality in the following 7 days. Other study, conducted by Zhang and Hong, presents a novel prediction score developed and validated specifically for patients with severe sepsis. Said model is based on Least Absolute Shrinkage and Selection Operator (LASSO) methodology and variables that are routinely used in clinical practice within the ICU grouped in the following categories: demographic data, laboratory variables, vital signs, comorbidities, vasopressors, Glasgow Coma scale and urine output. The LASSO score showed the best discrimination in the validation cohort as compared with other scores such as SAPS II, acute physiological score III (APS III), Logistic organ dysfunction system (LODS), SOFA, and OASIS.

In the hospital in general, important studies have been carried out in which exclusive models were developed for the prediction of mortality in patients with sepsis [66–68]. For the studies the cohort was not composed exclusively of ICU patients, and although some of the patients received ICU care, the selection criteria are fundamentally different from those of the other studies in which the patients were evaluated for sepsis at the time of admission to the ICU.

Lagu et al. developed a multilevel mixed-effects logistic regression model to predict in-hospital mortality in patients with sepsis using only administrative data; Predictors included patient demographics (age, sex, race, insurance type), site and source of sepsis, presence of 25 individual comorbidities, treatment (mechanical ventilation, vasopressors and admission to the intensive care unit). In the validation cohort, the model developed by Lagu et al. presented discriminatory ability statistically similar to traditional severity of disease scoring system, APACHE II, SAPS II and MPM III. The best performance on the validation cohort was obtained with the SAPS II score.

In 2014 Osborn et al. used the data from 23438 patients with suspected or confirmed sepsis from 218 hospitals in 18 countries to generate the Sepsis Severity Score. Even though the purpose of such a model is to predict in-hospital mortality for patients with sepsis during their ICU stay, not all the patients analyzed for its developed come from the Intensive care unit, moreover, the patient location at symptom onset is one of the predictors. Other predictors are the Geographic region, organ failure (Cardiovascular, Pulmonary, renal, hepatic, hematologic), conditions related to the vital signs (hyperglycemia, tachypnea, hypothermia, hyperthermia, hypotension), laboratory measures (Lactate, and white blood cell count), medicine intake (fluids and vasopressors) and treatments and conditions (mechanical ventilation, altered mental status and chills with rigor). The Sepsis Severity Score accurately estimated the probability of hospital mortality in severe sepsis and septic shock patients. It performed well with respect to calibration and discrimination.

Based on the results from Lagu et al, and Osborn et al; in 2016 Ford et al. developed a Severe Sepsis Mortality Prediction Model and Score that only used administrative data. Data from 108448 patients were used for the development of the mentioned score. Predictors included were demographics (age, gender and race), measures of acute illness severity (Mechanical ventilation, shock, hemodialysis and ICU care), and 20 comorbidities (anemia, depression, diabetes, drug and substance abuse, chronic lung disease, congestive heart failure, hypertension, hypothyroid disease, liver disease, lymphoma, metastatic carcinomas, neurologic conditions, obesity, paraplegia, perivascular conditions, psychiatric diseases, pulmonary circulatory, renal failure, malignant solid tumors, weight loss). The sepsis severity model and score presented an excellent discrimination performance and were well calibrated and far exceeded the performance obtained with the Osborn score. Table 2.6 presents an abridgment to the cited works.

2.6 Conclusion

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection; in recent years' clinicians have become aware of the importance of the long-term outcomes associated with this conditions. In Section 2.3 some studies were shown that indicate that Patients with sepsis have ongoing mortality beyond short-term end points, and survivors consistently demonstrate impaired quality of life. Even more, sepsis survivors suffer from additional morbidities such as higher risk of readmissions, cardiovascular disease, cognitive impairment and of death, for years following sepsis episode; despite this mortality prediction models for patients within the ICU focus on the short term mortality prediction (7-day mortality and in-hospital mortality).

All the works reported in Table 2.6 presented better performance than traditionally severity of disease scores, which supports the conclusion in chapter 1, and indicates that also for patients with sepsis the customize models perform better than the general population approaches; however, the use of in-hospital mortality as an end point for clinical studies are not enough to understand the effect of sepsis on mortality and quality of life, and the current understanding of the risk factors and mechanisms underlying long-term sequelae in patients that suffered from this condition is limited. Therefore, identify risk factors during an ICU stay that reverberate and even could predict long-term outcomes will help physicians offer better treatments.

Table 2.6 Sepsis severity of disease related works.

Author	year	Objective	Analysis tools	Outcome	Patients	Main result
Le Gall	1995	To develop customized versions of the Simplified Acute Physiology Score II (SAPS II) and the 24-hour Mortality Probability Model II (MPM II-24)	Logistic regression models	In-hospital mortality	1130	The customized models were well calibrated and outperform the discrimination obtained with the traditional scores.
Arabi	2003	To assess the in-hospital mortality prediction for four traditional severity of disease classification systems (APACHEII, SAPSII, MPMII-at admission, MPMII-after 24 hours) and the two customized models from [15].	Logistic regression models	In-hospital mortality	250	The overall mortality prediction was adequate for all six systems, however customized models showed improved calibration. Discrimination was best for customized MPM II-24.
Lagu	2011	To develop a sepsis model for the in-hospital mortality that uses only administrative data.	Logistic regression model	In-hospital mortality	166931 Not exclusively ICU patients according Angus Criteria. 36% received ICU care	The proposed sepsis mortality model has discrimination similar to and calibration superior to those of existing severity scores (APACHEII, MPMIII).
Osborn	2014	To develop a sepsis severity score that estimates the in-hospital mortality.	Logistic regression model	In-hospital mortality	23438 Not exclusively ICU patients	The Sepsis Severity Score accurately estimated the probability of hospital mortality in severe sepsis and septic shock patients. It performed well with respect to calibration and discrimination.
Carrara	2015	To develop a prediction model of 7-day mortality from vital signs and parameters routinely collected during the first 48 hours after septic shock onset	Linear regression model with regularization (elastic net)	7-day mortality	96	The developed model outperform the traditional scores for mortality risk assessment in ICU(SOFA, SAPSI)
Ford	2016	To develop, internally validate, and externally validate a severe sepsis mortality prediction model and associated mortality prediction score.	Logistic regression model	In-hospital mortality	563155 Not exclusively ICU patients according Angus Criteria. 35% received at least 1 day of ICU care	The model was well-calibrated and the AUC analysis support the models ability to discriminate patient mortality
Zhang	2017	To develop a score for in-hospital mortality prediction specifically for patients with sepsis	Least Absolute Shrinkage and Selection Operator (LASSO) regression	In-hospital mortality	3206	The LASSO score had good discrimination and calibration in a randomly selected subsample and outperform traditional severity of disease scores (SAPSII, APSIII, LODS, SOFA, OASIS, qSOFA)

CHAPTER 3. RESEARCH QUESTION

3.1 Introduction

Chapter 1 outlines how data gathered from patients in an intensive care unit is used in the form of scores, and also presents the most commonly used severity-of-illness classification systems. Traditional scores can be useful to guide prognostication, to assess ongoing disease development and organ function, to compare ICU performance over time and across units and to compare clinical trial population outcomes but they were not designed for individual prognostication. In order to increase the performance of outcome prediction scores and use them at the individual level, specific models according to groups of patients that shares a common characteristic have been developed.

In chapter 2 presents the modern concept of sepsis, the methods used for the clinical operationalization of the diagnosis and its incidence and hospital; it also describes several retrospective studies to identify sepsis by using different criteria that uses ICD codes and administrative data generated at the end of the hospital stay. Finally, it describes the long-term outcomes for sepsis and report studies focused on outcome prediction for sepsis patient within the ICU , which used in-hospital mortality as an end point.

In this chapter, the context presented in the previous chapters is used to structure the research question; the objectives and the outline of the methodology used in this thesis are also presented.

3.2 Research Question

In accordance with chapter 2, Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection [17] and patients who suffer it are often in a very delicate condition, with mortality rates around 26% [51]. Multiple models have been developed to predict outcomes exclusively of sepsis patients, which have proven to perform better that traditional severity-of-illness classification systems [62, 64–69]. However, these customize models are focused on the short-term mortality prediction. In recent years clinicians have become aware of the importance of the long-term outcomes associated with sepsis; only half of patients are alive at 1 year after a sepsis episode and the surviving patients are in risk of suffer other conditions like neurologic and respiratory impairments, renal failure, ischemic stroke, hemorrhagic stroke, myocardial infarction, heart failure, and sudden cardiac death or ventricular arrhythmia [56–58, 60, 61, 70].

Currently, the risk factors and mechanisms underlying long-term sequelae in patients that suffered from sepsis is limited, according to the above

3.3 Objectives

In order to answer the research question, the following objectives were proposed:

3.3.1 General objective

Generate, from demographic and physiological parameters, a methodology that stratifies the severity of commitment of a patient admitted to an intensive care unit, with sepsis; in order to initiate an early and appropriate treatment from their individual characteristics.

3.3.2 Specific objectives

1. Generate a segmentation that allows stratifying a patient into risk groups according to their characteristics of clinical relevance.
2. Develop a statistical model that relates the parameters of clinical relevance with the mortality of patients in each of the risk groups.
3. Evaluate the developed model, both from its statistical performance and its clinical usefulness.
4. Develop assistance software that allows clinicians to establish a risk score to understand the severity of each patient's condition and generate early alarms for the start of treatment.

3.4 Methodology

The activities developed to answer the research question were divided, as the Figure 3.1 shows, in four different stages.

The first stage is the customize mortality prediction modeling. It includes several steps: review the core literature, propose the input variables and to define the outcome of interest. The input variables proposed were chosen from the bibliographic revision, the criterion of the expert intensivists and the availability in the used database (which will be described in detail in the next chapter). At this stage, a first study cohort composed was selected and termed cohort A. Cohort A is composed of 5650 sepsis patients (explicit or according to Angus criteria) that had the majority of the variables of interest. The main outcomes of this stage are an ensemble model for the one-year mortality prediction of sepsis patients with in the ICU that outperformed traditional severity-of-illness scoring systems; and a subset of predictors that are truly related with one-year mortality. The subset of relevant variables was obtained using two methodologies: 1) Least Absolute Shrinkage and Selection Operator (LASSO); 2) Stochastic Gradient Boosting relative variable importance.

The second stage is the stratification of patients in risk groups. In this stage the definitive cohort (composed of 15082 admissions) was obtained, each admission included in the cohort meets the following characteristics:

- A retrospective sepsis diagnostic according to the explicit, Angus, Martin and Sepsis-3 criteria.
- Patients older than 16 years
- The majority of variables that were find to be relevant for the one-year mortality in the stage 1. This subset of predictors was complemented with variables that are frequently used within the ICUs.
- The Elixhauser comorbidity description (details of this will be presented in future chapters)

In this stage we generated two customized scoring systems for the assessment of the one-year mortality risk of sepsis patients within the ICU. The first score was based on dichotomization of the variables, and the second one used multiple cutoff points for each continuous numerical variable (i.e. the laboratory measurements, the routine charted data and the admission age).

The third stage is the personalized predictive modeling based on patient similarity, in which we use patient similarity metrics to identify a precision cohort for an index admission, this precision cohort is used to train a personalized model. For the construction of the precision cohort two parameters were adjusted: the modeling of the interaction between admissions and the number of similar admission used for the precision cohort. The number of admission used for the precision cohort was varied from 1000, to 13000;

and in order to model the relations between admission, five patient similarity measures were proposed and evaluated:

1. Cosine similarity (CS): in this approach the data of each admission is represented as an Euclidean feature vector, and the similarity of two admissions is computed as the angle between the two vectors.
2. Equally contribution similarity (ECS): In this approach a similarity term composed only by categorical data (like comorbidities, gender or treatments) is computed in an independent way and multiplied with the CS. The inclusion of such a term achieves that only the admissions that share a common characteristic are connected.
3. Weighted Contribution Similarity (WCS): This similarity measure is based on the idea that different conditions carry different mortality risk. For this reason, in this approach each variable of the categorical data is weighted. Three sets of weights were evaluated.

To determine the optimal combination of the mentioned parameters, personalized logistic regression models were generated for each pair of similarity measure and number of similar patients over a validation subset composed of 10% of the population. Personalized Stochastic Gradient Boosting (SGB) models were also generated with the parameters that presented the best performance.

In addition to the above, we evaluated a deep learning approach, which is composed of a multilayer neural network that is regularized by the patient similarity measure used for the personalized SGB models. In this stage our goals were: 1) Analyze the impact and relevance of the patient similarity metrics when patients are related by a common characteristic (a sepsis diagnosis) and a challenging outcome is evaluated (one-year mortality). 2) Evaluate if a single-time-trained non-linear method (a multi-layer neural network) that incorporates a similarity-based regularization could increase the prediction performance at the patient level and compete with a personalized model.

In the last stage we develop a software that could be used in the clinical environment with the models generated in the stages 2 and 3.

3.5 Significance of the research: Impact and products

The emergence of machine learning techniques in the field of health is a fact. Specifically, in the field of Intensive Care, it is undeniable that the potential for its application is immense; for instance, the development of a model that takes into account the peculiarities of sepsis and identify sensitively and early poor long-term patient's outcome, could become a very useful tool to help the clinical group to understand the severity of the disease and could help to the generation of alerts that favor early onset of therapeutic measures; thus helping to improve the prognosis of patients with sepsis admitted to an intensive care unit. Moreover, the identification of those patients who are at risk of dying one year after their sepsis-related ICU admission constitutes the first step in understanding the risk factors and mechanisms underlying long-term sequelae in patients that suffered from this condition.

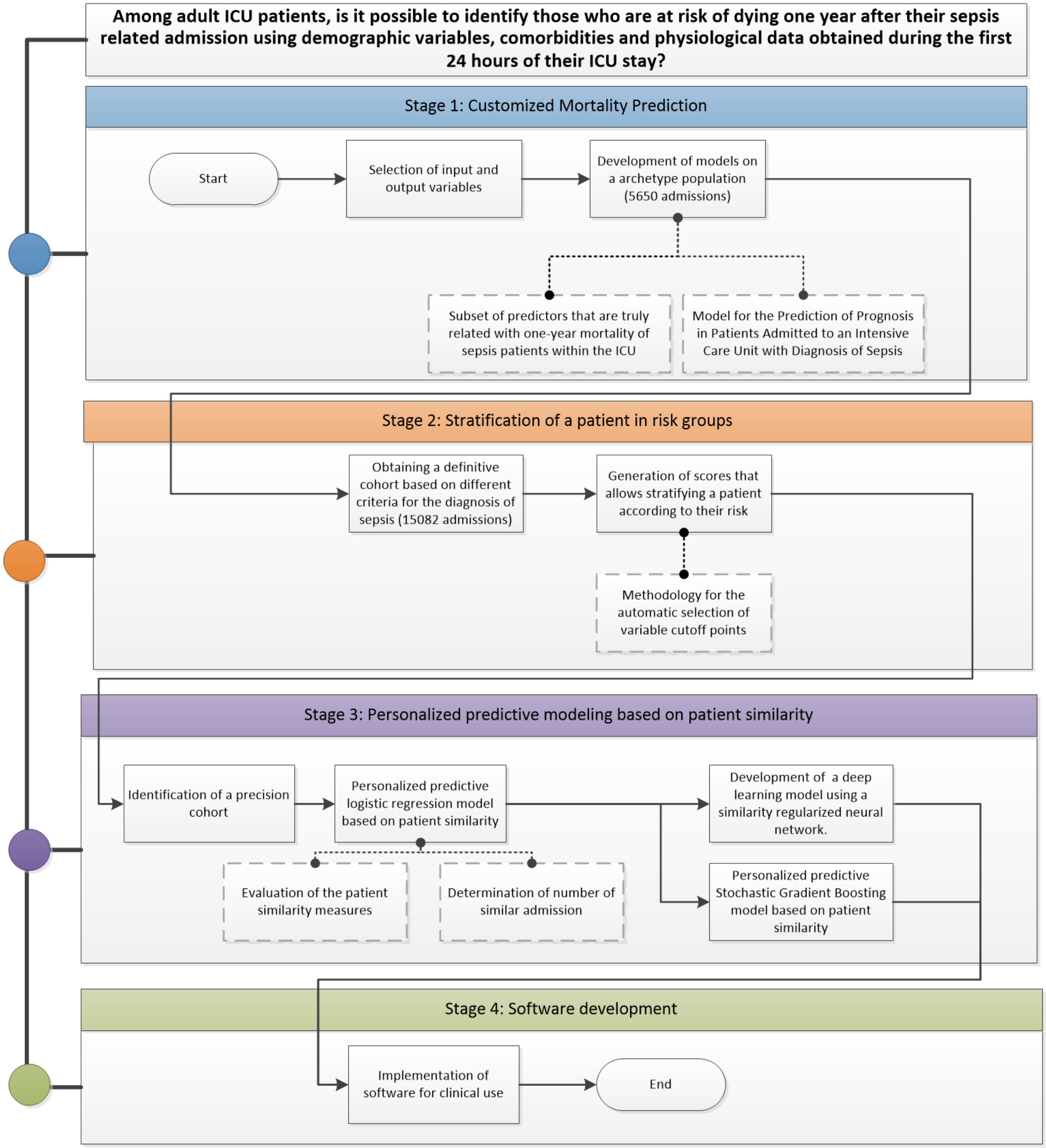

Among adult ICU patients, is it possible to identify those who are at risk of dying one year after their sepsis related admission using demographic variables, comorbidities and physiological data obtained during the first 24 hours of their ICU stay?
Stage 1: Customized Mortality Prediction
Start
Selection of input and output variables
Development of models on a archetype population (5650 admissions)
Subset of predictors that are truly related with one-year mortality of sepsis patients within the ICU
Model for the Prediction of Prognosis in Patients Admitted to an Intensive Care Unit with Diagnosis of Sepsis
Stage 2: Stratification of a patient in risk groups
Obtaining a definitive cohort based on different criteria for the diagnosis of sepsis (15082 admissions)
Generation of scores that allows stratifying a patient according to their risk
Methodology for the automatic selection of variable cutoff points
Stage 3: Personalized predictive modeling based on patient similarity
Identification of a precision cohort
Personalized predictive logistic regression model based on patient similarity
Development of a deep learning model using a similarity regularized neural network.
Evaluation of the patient similarity measures
Determination of number of similar admission
Personalized predictive Stochastic Gradient Boosting model based on patient similarity
Stage 4: Software development
Implementation of software for clinical use
End

Figure 3.1. Methodology.

It is also expected that the developed tool will support intensivists in decision making within an intensive care unit; since the proposed model will allow the development of software that indicates which demographic variables, comorbidities and physiological data are relevant to the one-year mortality of each particular sepsis patient. We also consider that this project would help the strengthening of the scientific community and the generation of new knowledge, because, currently, there is no methodology for the identification the long-term mortality in ICU patients with a sepsis diagnostics and there are no studies that indicate the usefulness of similarity metrics in cases in which patients have a common characteristic, such as the diagnosis.

PART 2: STUDY COHORT

In the first part of this thesis we indicated that we are interesting in identify patients who are at risk of dying one year after their sepsis related admission. Part 2 focuses in how we obtained an appropriate study cohort. Chapter 4 presents the database from which we take the admissions that compose our study cohort, called Medical Information Mart for Intensive Care (MIMIC); it also provides a list of existing clinical databases already in use for research. Chapter 5 Presents the details of the study cohorts that we selected, such as the admission inclusion criteria, and the variables used for subsequent analysis.

CHAPTER 4. CLINICAL DATABASE

4.1 Introduction

As stated in chapter 1, the critical condition of an intensive care unit patient requires close and constant monitoring, which generates a large volume of data that can be used for the development and evaluation of applications, systems and models based on computational tools, such as machine learning-aided medical software. Despite the obvious usefulness of patient driven data accumulated with clear structure that make it meaningful and usable, currently in Colombia, many hospitals do not have established databases that archived and organized detailed patient data into central repositories, and those who have them do not make efficient use of their data. On the other hand, meaningful clinical conclusions can only be obtained with a sufficiently representative sample to generalize the results.

In this chapter we describe in detail the used database, MIMIC-III (Medical Information Mart for Intensive Care), a large, single-center database comprising information relating to patients admitted to critical care units at a large tertiary care hospital, and we also present other databases that are readily accessible.

4.2 The Medical Information Mart for Intensive Care (MIMIC) Database

MIMIC dataset (Multi-parameter Intelligent Monitoring for Intensive Care) is widely used for researchers, evaluators of algorithms and physiological data analysis systems in the ICU. Its first version includes data from 100 patients, each record has between 24 and 48 hours of data recorded from patient monitors (Electrocardiogram, blood pressure, respiration and oxygen saturation) accompanied by detailed clinical data derived from the patient's medical record and notes during monitoring, which provides the context of each patient condition [71].Figure 4.1 shows a 3-hour extract of several physiological measurements, an alarm is highlighted in red around the systolic blood pressure.

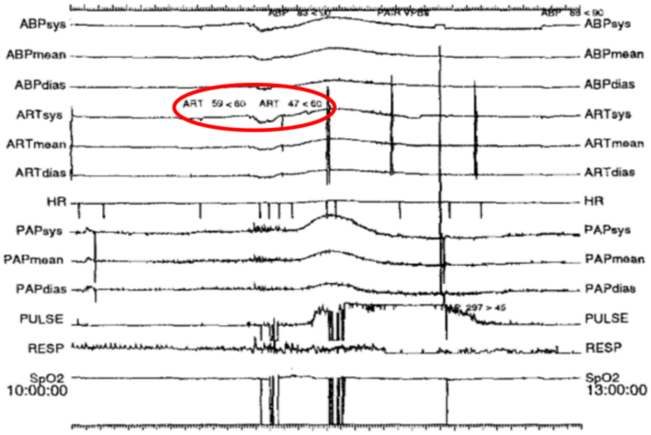


Figure 4.1. Excerpt from a MIMIC record. Taken from [71].

MIMIC includes patients that were believed to be hemodynamically unstable; therefore, they are considered to adequately represent the range of pathologies that result in abrupt changes in blood pressure. MIMIC is of particular interest in research regarding heart rate, blood pressure, respiratory dynamics and their interactions [71]. However, MIMIC patients do not represent the entire population in

ICU; as result, a larger bank of information became necessary in order to develop and evaluate systems that could assist ICU clinical staff in decision-making and the outcome prediction,.

A new database version MIMIC-II [72–74] was created to take advantage of technological advances in telecommunications and storage systems. It has more than 20,000 patients and in addition to the clinical history and physiological variables, includes laboratory data, therapeutic interventions, progress notes, radiology reports and ICD9 diagnostic codes. MIMIC-II is a great research resource due to four factors:

- It is open, since it allows researchers from all over the world to access it free of charge after they request access.
- It has data with high temporal resolution.
- It keeps patient confidentiality.

MIMIC-II dataset were collected in four intensive care units: medical (MICU), surgical (SICU), coronary care unit (CCU), and cardiac surgery recovery unit (CSRU) at Beth Israel Deaconess Medical Center in Boston, MA, USA during the period from 2001 to 2008. MIMIC-II records were deidentified by removing protected health information. Also, all hospital admissions and ICU stays of each patient were time-shifted to a hypothetical period in the future [72].

MIMIC-II consists of two main components, clinical data and physiological signals. Clinical data included demographic information, drug doses, nursing notes, discharge summaries, nurse verified hourly vital signs and laboratory test results, and were organized in a relational database, Table 4.1 describes different clinical data types in MIMIC-II by giving examples of each type. The signals, which include the continuous records of vital signs, were stored in an open format, which makes it possible to read them under any operating system. Figure 4.2 presents an extract of the waveform signals that can be found in MIMIC-II.

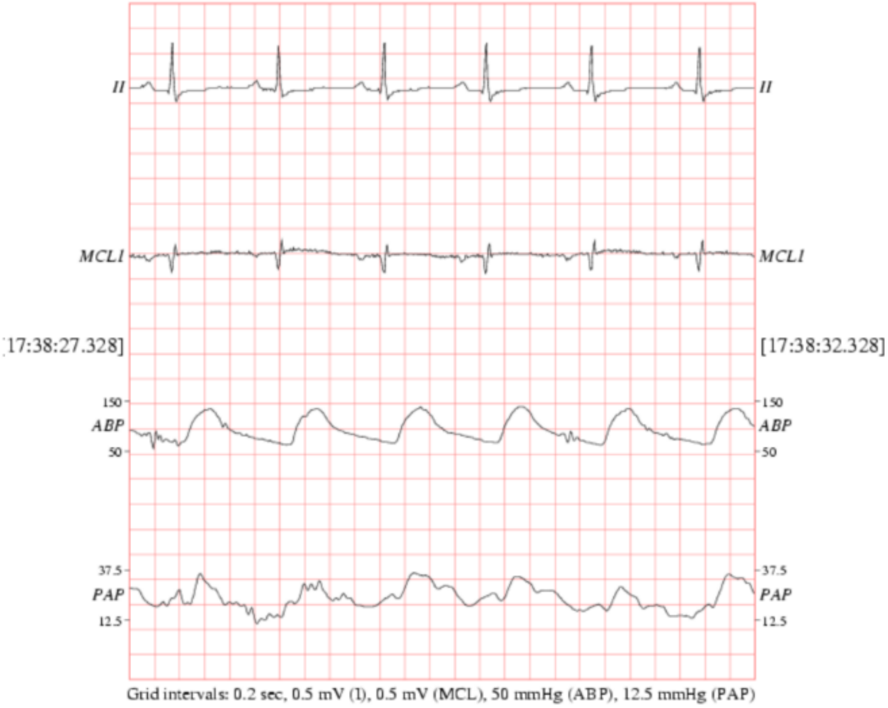

II
II
MCLI
MCLI
17:38:27.328]
[17:38:32.328]
150
ABP
50
150
ABP
50
37.5
PAP
12.5
37.5
PAP
12.5
Grid intervals: 0.2 sec, 0.5 mV (I), 0.5 mV (MCL), 50 mmHg (ABP), 12.5 mmHg (PAP)

Figure 4.2. Waveforms signals from MIMIC-II. Taken from [72].

Table 4.1. Clinical data types in MIMIC-II

Clinical Data Type	Examples
Demographics	Age, gender, date of death, ethnicity
Hospital admission	Admission and discharge dates, ICD-9 codes
Free-text	Reports of imaging studies and 12-lead ECGs, nursing notes
Intervention	Ventilator settings, Intravenous medications
Laboratory tests	Blood chemistries, hematology, urinalysis, microbiologies
Severity scores	SAPS I, SOFA, Elixhauser comorbidities
Fluid balance	Solutions, blood transfusion, urine output, estimated blood loss

MIMIC-II includes 26.870 adult hospital admissions and 31.782 adult ICU stays. MICU patients constitute the largest group among the four care units. The overall median ICU and hospital lengths of stay were 2.1 and 7 days, respectively. The overall in-hospital mortality was 11.5%, however the mortality of the CSRU patients was very low. Table 4.2 presents some MIMIC-II statistics for the adult population stratified with respect to the four critical care units.

Table 4.2. Adult patient statistics in MIMIC-II

	MICU	SICU	CSRU	CCU	Total
Hospital admissions	10313 (38,4%)	6925 (25,8%)	5691 (21,2%)	3941 (14,7%)	26870 (100%)
Distinct ICU stays	12648 (39,8%)	8141 (25,6%)	6367 (20,0%)	4626 (14,6%)	31782 (100%)
Age	64,5	61,1	67,1	71,4	65,5
Gender (Male)	6301 (49.8%)	4701 (57.7%)	4147 (65.1%)	2708 (58,5%)	17857 (56.2%)
ICU length of stay (days)	2.1	2.4	2.1	1.9	2.1
Hospital length of stay (days)	7	8	8	5	7
Hospital mortality	1645 (16.0%)	842 (12.2%)	213 (3,7%)	392 (10.0%)	3092 (11,5%)

MIMIC-II has been used for a large number of analytical studies, which include epidemiological studies, development of clinical decision rules, reducing false alarm within the ICU and prediction of important physiological values and adverse events [75–79]. Although MIMIC II is innovative and unprecedented, it still has some limitations. The administration of oral medications is not automatically entered, but is part of the nursing notes. The data are exclusively from the stay in the ICU, so sometimes the context that would provide knowledge of the history prior to admission to the intensive care unit is missing. The information in the database reflects the actual protocols of the Beth Israel medical center, so it is possible that researchers from other institutions do not find in MIMIC II information regarding specific procedures that are not performed in the hospital where the data was collected.

In order to solve some of these limitations, in November 2015 MIMIC-III [80] was launched, whose name changed from "Multi-parameter Intelligent Monitoring for Intensive Care" to "Medical Information Mart for Intensive Care". This new version provides demographic information, vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and mortality (both inside and outside the medical center).

The data of MIMIC-III was obtained from two sources, one is external, the Social Security Administration Death Master File which is used to obtain the out-of-hospital mortality dates; and the other is from the hospital's information systems. Figure 4.3 presents an overview of the MIMIC-III database. The hospital data comprises seven important blocks:

- Bedside monitoring that includes vital signs, trends, alarms and waveforms; although only a subset of patient records include physiological waveforms obtained from bedside monitors (such as electrocardiograms, blood pressure waveforms, photoplethysmograms, impedance pneumograms);
- Chart data that includes fluids balance, medicine administration and progress notes
- Laboratory and microbiology tests.
- Billing that include disease, drugs and procedures codes.
- Demographics that include admission and discharge dates, dates of birth and death, religion, ethnicity and marital status.
- Notes and reports that includes discharge summaries and imaging reports.

MIMIC-III records was deidentified using structured data cleansing and date shifting. The deidentification process removes or changes fields such as patient name, telephone number, address, and dates. In particular, stay date fields were shifted into the future by a random offset for each individual patient in a consistent manner in order to preserve intervals, resulting in stays which occur sometime between the years 2100 and 2200. Dates of birth for patients aged over 89 were shifted to obscure their true age, these patients appear in the database with ages of over 300 years.

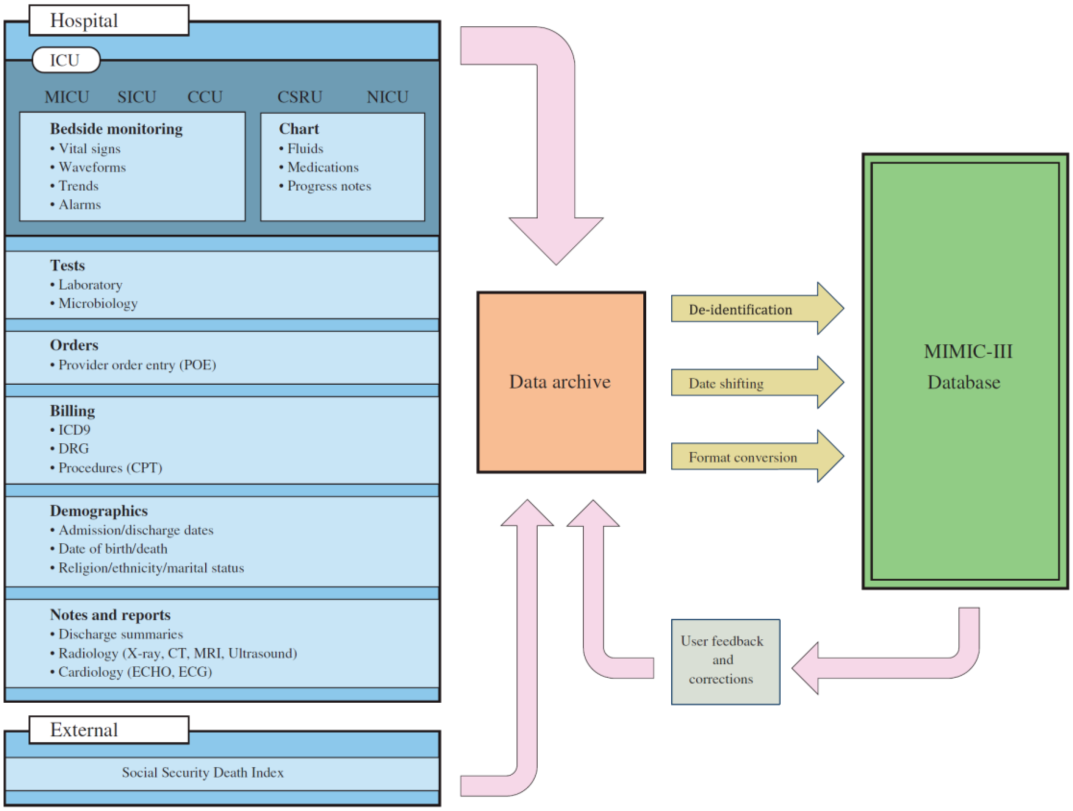


Figure 4.3. Overview of the MIMIC-III critical care database. Taken from [80].

The description of the adult population is presented in Table 4.3. MIMIC III contains data associated with 53,423 different ICU stays from 49,785 hospital admissions for 38,597 distinct patients older than 16 years who entered the ICU between 2001 and 2012. The median age for adult patients is 65.8 years, 55.9% are male and in-hospital mortality rate is 11.5%. The median length of stay in the ICU is 2.1 days, and the median length of hospital stay is 6.9 days. A mean of 4579 charted observations and 380 laboratory measurements are available for each hospital admission. The second most common International Classification of Diseases code for patients aged 16 years and above was 038.9 (*'Unspecified septicemia'*), accounting for 4.2% of all hospital admissions [80].

Table 4.3. Details of the MIMIC-III adult patient population.

	MICU	SICU	CCU	CSRU	TSICU	Total
Hospital admissions	19770 (39,7%)	8110 (16,3%)	7258 (14,6%)	9156 (18,4%)	5491 (11,0%)	49785 (100%)
Distinct ICU stays	21087 (39,5%)	8891 (16,6%)	7726 (14,5%)	9854 (18,4%)	5865 (11.0%)	53423 (100%)
Age	64,9	63,6	70,1	67,6	59,9	65,8
Gender(Male)	10193 (51,6%)	4251 (52,4%)	4203 (57,9%)	6000 (65,5%)	3336 (60,7%)	27983 (55.9%)
ICU length of stay (days)	2,1	2,3	2,2	2,2	2,1	2,1
Hospital length of stay (days)	6,4	7,9	5,8	7,4	7,4	6,9
Hospital mortality	2859 (14,5%)	1020 (12.6%)	817 (11,3%)	424 (4,6%)	628 (11,4%)	5748 (11,5%)

4.3 Other medical databases

4.3.1 PCORnet

PCORnet [81], the National Patient-Centered Clinical Research Network, is a large network that collects data routinely gathered in a variety of healthcare settings, including hospitals, doctors' offices, and community clinics. PCORnet objective is to empower individuals and organizations to use data to answer practical questions that help patients, clinicians, and other stakeholders to make informed healthcare decisions [82].

PCORnet is a Distributed Research Network that captures clinical data and health information that are created every day during routine patient visits. In addition, PCORNet is using data shared by individuals through personal health records or community networks with other patients as they manage their conditions in their daily lives.

Currently, PCORnet represents more than 100 health institutions across the United States and have data on more than 100 million Americans. Data from all of these patients are potentially available for observational research.

4.3.2 NHS Open data

The National Health Services (NHS England) is a governmental entity that retains one of the largest repositories of data on people's health in the world. One of NHS England projects is Open data a publicly released information, often from the government or other public bodies, which is made freely available to everyone to use, its main objective is to increase transparency and trace the outcomes and efficiency of the British healthcare sector. An example of the use of Open data is LG Inform, an application which pulls together disparate data about local services, including open data from UK local authorities, and provides users with meaningful information about a local area and allows users to review and compare services between authorities using charts, tables and maps [82].

4.3.3 The eICU Collaborative Research Database

The eICU Collaborative Research Database (eICU) [83] is a multi-center intensive care unit database with high granularity data for comprises 200,859 patient unit encounters for 139,367 unique patients admitted between 2014 and 2015 to one of 335 units at 208 hospitals located throughout the United States. eICU was made available to the public in 2018 and include vital signs, laboratory measurements, medications, care plan information, admission diagnosis, patient history, time-stamped diagnoses and similarly chosen treatments. The data are organized into tables which broadly correspond to the type of data contained within the table. The eICU database resulted from the alliance of Philips Healthcare which provides a teleICU, a centralized model of care where remote providers monitor ICU patients continuously, providing both structured consultations and reactive alerts, service known as the eICU program; and the Laboratory for Computational Physiology at MIT which has previously shared MIMIC database.

4.3.4 Biologic Specimen and Data Repositories Information Coordinating Center

The National Heart, Lung, and Blood Institute (NHLBI) is an US National Institute of Health that provides global leadership in the prevention and treatment of heart, lung, and blood diseases and supports basic, translational and clinical research in these areas. In 2008, the NHLBI established the Biologic Specimen and Data Repositories Information Coordinating Center (BioLINCC) to expand the utilization of two unique research resources developed and maintained by the NHLBI: the NHLBI Biologic Specimen Repository (Biorepository) and the NHLBI Data Repository. Many of the clinical studies in the Data Repository have associated biospecimen collections stored in Biorepository [84].

The main objectives behind the BioLINCC is to maximize the scientific value of the Biorepository and Data Repositories, and to promote the availability and use of other NHLBI-funded population-based biospecimen and data resources. The mission of the NHLBI Biorepository is to acquire, store and distribute quality biospecimens to the scientific community using standardized processes and procedures approved by the NHLBI, the Biorepository has several plasma, serum and whole blood collections from epidemiologic studies conducted in blood donors and transfusion-recipients. Research on these biospecimens enabled key advancements in transfusion safety including evaluation of donor screening assays for viral agents such as HIV, hepatitis B and hepatitis C, and risk estimations for transfusion-transmitted viral agents. The NHLBI has supported data collection from participants in epidemiology studies and clinical trials for over six decades. These data have often been sent to the NHLBI at the conclusion of the study and placed in the Data Repository. The Data Repository is managed by the Epidemiology Branch in the Division of Cardiovascular Sciences and includes individual level data on hundreds of thousands of participants from 200 Institute-supported clinical trials and observational studies [85].

Within BioLINCC there are three studies with both resources Specimens and Datasets that are partially related to sepsis; the first one focus on determine if dietary supplementation of omega-3 fatty acids, γ-linolenic acid and antioxidants to patients with early acute lung injury or sepsis-induced respiratory failure would increase ventilator-free days; the second one was intended to assess the efficacy and safety of oral rosuvastatin in patients with sepsis-induced Acute Lung Injury and test the hypothesis that rosuvastatin therapy would improve the clinical outcomes of critically ill patients with sepsis-associated acute respiratory distress syndrome, both of this studies are framed into the Acute Respiratory Distress Network (ARDSNet), a randomized controlled trial conducted from 1996 through 2006. The third study aims to determine whether or not treatment with hydroxyurea titrated to maximum tolerated doses would reduce the frequency of vaso-occlusive crises by at least 50% in this case sepsis is one of the reportable events.

4.3.5 Intensive Care National Audit & Research Centre Case Mix Programme

The Intensive Care National Audit & Research Centre (ICNARC) is an independent charity established in 1994. The Case MixProgramme (CMP) is a national, comparative audit of patient outcomes from adult, general critical care units in England, Wales and Northern Ireland coordinated by the ICNARC. Data collected for the CMP take the following forms [86]:

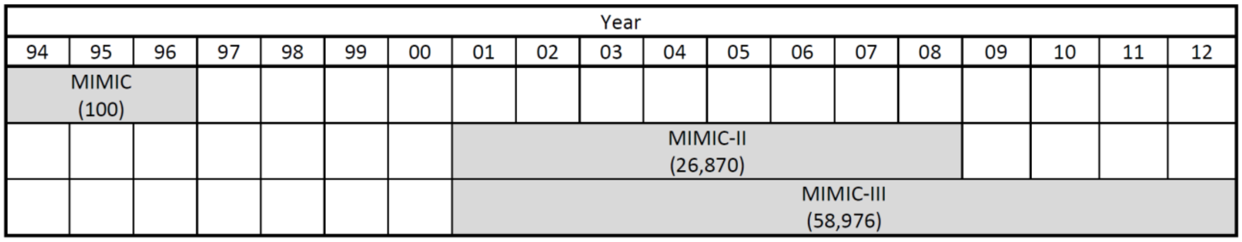
- Patient identifiers: admissions are identified by an admission number and an alphanumeric unit code
- Demographics: date of birth, gender and postcode
- Stay variables: Raw physiological data are collected to enable calculation of the APACHE-II, APACHE-III, SAPS-II and MPM-II scores and hospital mortality probabilities. Both the lowest and highest recorded values during the first 24 hours in the CMP unit are collected.
- Outcome: Survival data are recorded at discharge from the unit and from the hospital.
- Activity: length of stay in the ICU and in the hospital are recorded along with information about transfers between units or hospitals.

The vast amount of data have been used to produce numerous local and national analyses, specifically, there are four reports of analysis on the Case Mix Programme database focused on sepsis. The first one studies the admissions with neutropenic sepsis in adult, general critical care units between the April 1 of 2007 to September 30 of 2010. The second one studies the admissions with severe sepsis in adult, general critical care units from the January 1 of 2008 to the December 31 of 2009. The third one reports the number of sepsis admissions to critical care and their mortality between the April 1 of 2010 to March 31 of 2013. The final one presents the length of stay, survival and organ support of admissions with septic shock in 2012.

4.4 Conclusion

High-quality clinical databases are of value in clinical practice, in managing services and in developing health technologies. The use of inappropriate, unrepresentative or poor-quality data can lead to imprecise and inaccurate conclusions; for this reason, selection of the optimal database for a particular question is a crucial part of relevant analyses.

The selected database for this work is MIMIC-III which data comes from a single institution (Beth Israel Deaconess Medical Center in Boston Massachusetts). However, despite the limitation of being single-centered, the main advantages of MIMIC-III are: (1) Right now the only freely accessible critical care database of its kind. (2) the dataset spans more than a decade, Figure 4.4. (3) It has detailed information about individual patient care that includes time-stamped nurse-verified physiological measurements and out-of-hospital mortality dates. For this reasons MIMIC-III (and specially it previous version MIMIC-II) are widely used internationally.



<table>
<tr><th colspan="19">Year</th></tr>
<tr><td>94</td><td>95</td><td>96</td><td>97</td><td>98</td><td>99</td><td>00</td><td>01</td><td>02</td><td>03</td><td>04</td><td>05</td><td>06</td><td>07</td><td>08</td><td>09</td><td>10</td><td>11</td><td>12</td></tr>
<tr><td colspan="3">MIMIC
(100)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="8">MIMIC-II
(26,870)</td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="12">MIMIC-III
(58,976)</td></tr>
</table>

Figure 4.4. Period in years in which the admissions of the different versions of MIMIC were taken.

Within the other medical databases described before three contains intensive care data BioLINCC, CMP and eICU. BioLINCC data comes from randomized controlled trials and prospective cohort studies for this reason data presents considerable clinical detail and includes clinical physiology, severity of illness, and patient outcomes, however, this data are collected specifically for the purposes of answering a research question, this means that for each study only the variables and outcomes related to the question are recorded, in particular the BioLINCC studies that focus on sepsis includes data of a subset of patients with respiratory dysfunction and do not record long-term mortality. CMP only registers the physiological data associated with four severity-of-illness classification systems, moreover, only the lowest and highest values during the first 24 hours are recorded, which means that other statistical descriptors like the mean could not be used in works with this database; on the other hand, CMP do not contains outcomes beyond ultimate hospital discharge.

The eICU dataset is the most like MIMIC and its main advantage if the fact that it is multi-centered; however, some features of this database makes it not suitable for this project. eICU database is sourced from the eICU Telehealth Program (a model that allow caregivers from remote locations to monitor treatments for patients, alert local providers to sudden deterioration, and supplement care plans) which makes the amount of data obtained from each patient less than the amount of data that is obtained within the Beth Israel Deaconess Medical Center ICUs, this is evidenced, for example, in the fact that for the eICU database a total of 158 distinct types of laboratory measurements are captured and for MIMIC-III 753 laboratory measures were recorded. Second, and most important, eICU reports the health status at hospital discharge but do not include out-of-hospital mortality dates and for this study in particular, an important advantage of MIMICIII is that besides in-hospital mortality, MIMIC-III provides mortality dates through the Social Security Administration Death Master File.

CHAPTER 5. STUDY COHORT

5.1 Introduction

This chapter describes how we obtain the admissions of MIMIC-III that present a diagnosis of sepsis from the identification criteria presented in chapter 2. After that, we detail the exclusion criteria and the variables taken into account for the creation of two different study cohorts. At the end of the chapter, we present a description of the population of the study cohorts.

5.2 Retrospective sepsis identification in MIMIC-III

Chapter 2 presents four methods of retrospective identification of sepsis, this section will describe how these criteria are implemented within MIMIC-III.

Figure 5.1 shows the ICD-9 code structure. It has an alphabetic or numeric as first digit, whereas the remaining digits are numeric, the complete code has a minimum of three digits and a maximum of five digits. The first three digits indicate the category; the following decimals represent the cause, the origin, the anatomical site or the manifestations. We performed a preprocessing of the MIMIC-III ICD-9 codes in order to extract extracts the characters that represent the broad categories of diagnosis (when the first three digits are searched for) or specific diagnosis or procedures (when four or five digits are searched for).

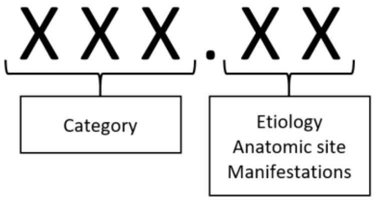
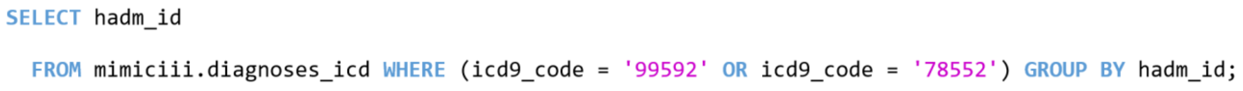


Figure 5.1. ICD-9 code structure

On the other hand, the codes for extracting admissions based on the explicit sepsis, Angus criteria and Martin criteria were build up from the "*diagnoses_icd*" and "*procedures_icd*" MIMIC-III tables; this tables contains the International Classification of Diseases Version 9 (ICD-9) codes for diagnoses and procedures for each hospital admission.

5.2.1 Explicit sepsis

In MIMIC-III, the identification of admissions that presents explicit sepsis is done using the "*diagnoses_icd*" table, which contains ICD diagnoses for patients, specifically ICD-9 diagnoses. In this table either of two ICD-9 diagnosis codes are looked for: 995.92 for severe sepsis or 785.52 for septic shock. Figure 5.2 presents the SQL query that was used to identify the explicit sepsis patients; it brings back 4085 unique hospital admissions (field hadm_id).

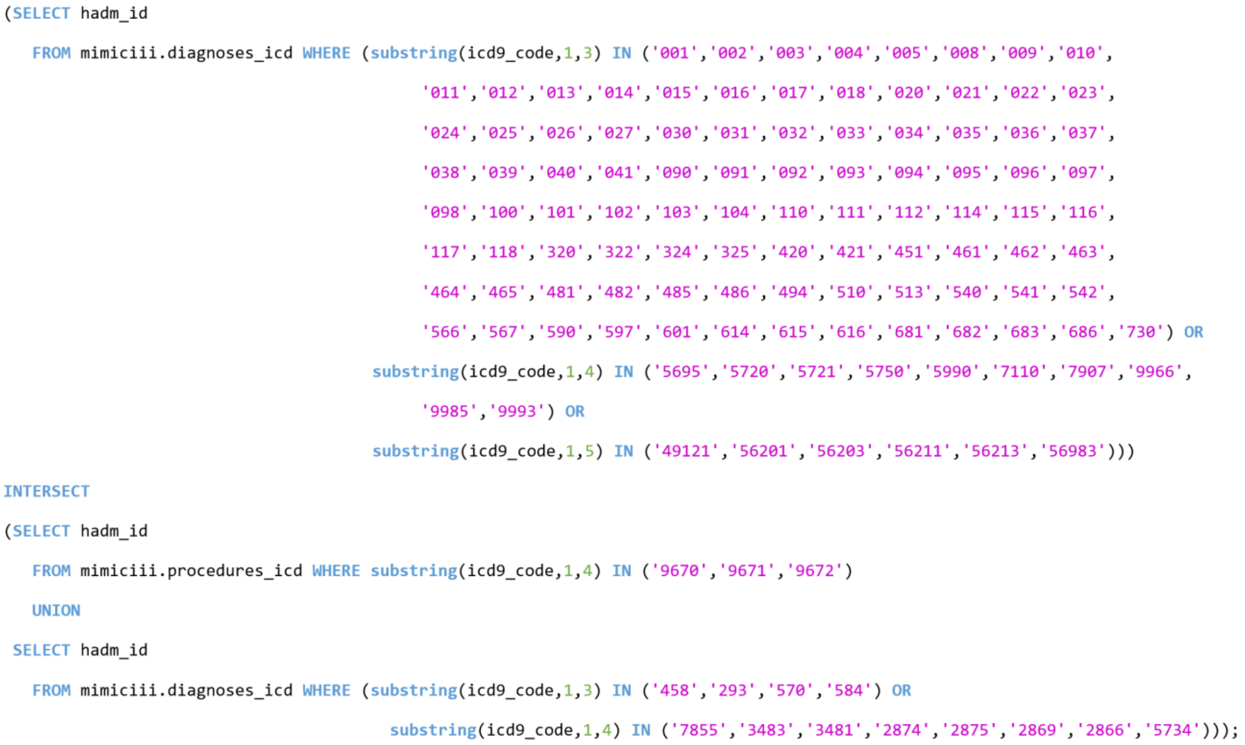


```
SELECT hadm_id
  FROM mimiciii.diagnoses_icd WHERE (icd9_code = '99592' OR icd9_code = '78552') GROUP BY hadm_id;
```

Figure 5.2. Query to obtain the explicit sepsis admissions.

5.2.2 Angus criteria

For the identification of sepsis admissions using the Angus criteria [47] in MIMIC-III, We look for hospital admissions that presents at least one of 2097 bacterial or fungal Infection related ICD-9 codes ("diagnoses_icd" table). Additionally, we identify the hospital admissions that presents any organ dysfunction; i.e , any of 37 unique ICD-9 codes based on the classification of acute organ dysfunction codes presented in chapter 2 for the Angus criteria (34 related to diagnosis and 3 related to mechanical ventilation procedure). The admissions that present a sepsis episode according to the Angus criteria are those that present both infection and organ dysfunction. Figure 5.3 presents the SQL query that was used to identify the Angus criteria sepsis patients, it brings back 15.149 unique hospital admissions.

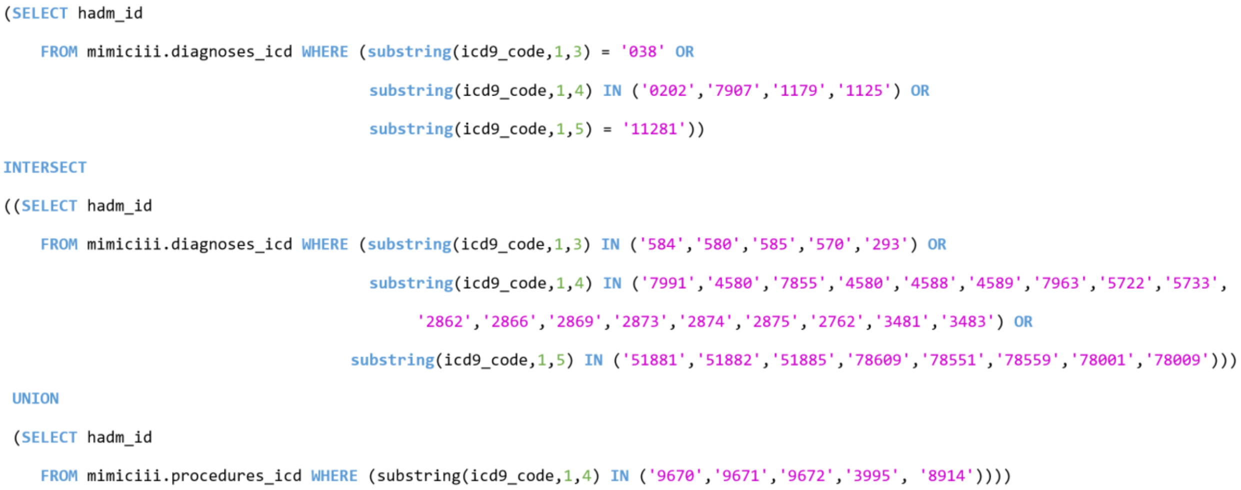


```
(SELECT hadm_id
   FROM mimiciii.diagnoses_icd WHERE (substring(icd9_code,1,3) IN ('001','002','003','004','005','008','009','010',
                                          '011','012','013','014','015','016','017','018','020','021','022','023',
                                          '024','025','026','027','030','031','032','033','034','035','036','037',
                                          '038','039','040','041','090','091','092','093','094','095','096','097',
                                          '098','100','101','102','103','104','110','111','112','114','115','116',
                                          '117','118','320','322','324','325','420','421','451','461','462','463',
                                          '464','465','481','482','485','486','494','510','513','540','541','542',
                                          '566','567','590','597','601','614','615','616','681','682','683','686','730') OR
                                     substring(icd9_code,1,4) IN ('5695','5720','5721','5750','5990','7110','7907','9966',
                                          '9985','9993') OR
                                     substring(icd9_code,1,5) IN ('49121','56201','56203','56211','56213','56983')))
INTERSECT
(SELECT hadm_id
   FROM mimiciii.procedures_icd WHERE substring(icd9_code,1,4) IN ('9670','9671','9672')
   UNION
 SELECT hadm_id
   FROM mimiciii.diagnoses_icd WHERE (substring(icd9_code,1,3) IN ('458','293','570','584') OR
                                       substring(icd9_code,1,4) IN ('7855','3483','3481','2874','2875','2869','2866','5734')));
```

Figure 5.3. Query to obtain the sepsis admissions according to the Angus criteria. The first SELECT statement retrieves the admission with infection, the second SELECT statement brings back the admissions in which the mechanical ventilation procedure was presented, the third SELECT statement retrieves the admissions that present any diagnosis associated with the Angus acute organ dysfunction codes. The union of the second and third SELECT statements are intersected with the first SELECT statement to obtain the hospital admissions that have sepsis according to the Angus criteria.

5.2.3 Martin Criteria

For the identification of sepsis admissions using the Martin criteria [48] in MIMIC-III, first we search for septicemia, septicemic, bacteremia, disseminated fungal infection, disseminated candida infection, septicemic plague or disseminated fungal endocarditis codes in the "*diagnoses_icd*" table. To do that, we obtain hospital admissions that presents at least one of 20 different sepsis related ICD-9 codes. Then, we identify the hospital admissions that presents any organ dysfunction, for this we obtain the hospital admissions that presents any of 63 unique ICD-9 codes based on the classification of acute organ dysfunction codes presented in chapter 2 for the Martin criteria (58 related to diagnosis and five related to mechanical ventilation, hemodialysis and electroencephalogram procedures). The admissions that present a sepsis episode according to the Martin criteria are those that present both sepsis and organ dysfunction. Figure 5.4 presents the SQL query that was used to identify the Martin criteria sepsis patients, it brings back 6931 unique hospital admissions.



```
(SELECT hadm_id
    FROM mimiciii.diagnoses_icd WHERE (substring(icd9_code,1,3) = '038' OR
                                       substring(icd9_code,1,4) IN ('0202','7907','1179','1125') OR
                                       substring(icd9_code,1,5) = '11281'))
INTERSECT
((SELECT hadm_id
    FROM mimiciii.diagnoses_icd WHERE (substring(icd9_code,1,3) IN ('584','580','585','570','293') OR
                                       substring(icd9_code,1,4) IN ('7991','4580','7855','4580','4588','4589','7963','5722','5733',
                                             '2862','2866','2869','2873','2874','2875','2762','3481','3483') OR
                                     substring(icd9_code,1,5) IN ('51881','51882','51885','78609','78551','78559','78001','78009')))
 UNION
 (SELECT hadm_id
    FROM mimiciii.procedures_icd WHERE (substring(icd9_code,1,4) IN ('9670','9671','9672','3995', '8914'))))
```

Figure 5.4. Query to obtain the sepsis admissions according to the Martin criteria. The first SELECT statement retrieves the admission with sepsis, the second SELECT statement retrieves the admissions that present any diagnosis associated with the Martin acute organ dysfunction codes. The third SELECT statement brings back the admissions in which the mechanical ventilation, hemodialysis or electroencephalogram procedures were implemented. The union of the second and third SELECT statements are intersected with the first SELECT statement to obtain the hospital admissions that have sepsis according to the Martin criteria.

5.2.4 Sepsis-3

The identification of Sepsis-3 admission was made in [1] by Desautels et al, who shared with us the 2577 Sepsis-3 admissions. In their work, they take the earliest culture draw or antibiotic administration as the time of suspicion of infection, and then they define a window of up to 48 hours before this time and 24 hours after this time; The SOFA score at the beginning of this window was compared with its hourly value throughout this window, and when the hourly value was ≥ 2 points higher than the value at the start of the window the particular admission was designate as septic.

5.3 Study cohorts

From the identified sepsis patients, two different study cohorts were gathered. The first one (cohort A) was used to generate a customize one-year mortality prediction model of patients with sepsis admitted to an ICU; for this cohort, by recommendation of the experts, we use the Angus and the explicit sepsis criteria. The Angus implementation of the 2001 international consensus conference definition of severe sepsis offers a reasonable approach to identifying patients with severe sepsis, it is a validated protocol and of the four criteria used in this work is the one that returns a greater number of admissions. On the other hand, the explicit sepsis criterion is the only one that is based on the clinical judgment of an expert, since it indicates that a treating doctor considered that a patient had severe sepsis or septic shock during his/her hospitalization.

The second cohort (cohort B) was used to developed both a system for the stratification of patients in risk groups and a method for the personalized predictive modeling based on patient similarity. For the second of these tasks it was necessary to have a large cohort in order to increase the possibility of finding similar admissions. According to the above, this cohort was selected from all methodologies for identifying patients with sepsis.

5.3.1 Cohort A

It was selected from the 58977 MIMIC-III admissions, all the ones that complying with the following: i) ICD-9-CM codes for both a bacterial or fungal infections and a diagnosis of acute organ dysfunction according to the Angus criterion ii) Explicit sepsis related diagnosis: severe sepsis and septic shock. 15254 admissions. Table 5.1 depicts predictor variables for cohort A.

Table 5.1. Extracted predictors for cohort A.

Parameter	Unit	Parameter	Unit
LABORATORY MEASUREMENTS		COMORBIDITIES	
Platelet Count	10^9/L	Diabetes	Binary (Presence)
Bilirubin	mg/dL	Immunosuppressive diseases	
Creatinine	mg/dL	Malignancy	
Fraction of Inspired Oxygen (FiO2)	%	Hematologic malignancy	
Partial pressure arterial oxygen and fraction of inspired oxygen ratio (PaO2/FiO2)	Ratio	Metastatic cancer	
White Blood Cell (WBC) count	10^3/mm^3	Heart failure	
Potassium	mEq/L	Pulmonary diseases	
Sodium	mEq/L	Vascular diseases	
Bicarbonate	mEq/L	Coronary diseases	
Lactate	mg/dL	Obesity	
Arterial pH	pH	Alcohol abuse	
Hematocrit	%	Collagen diseases	
Hemoglobin	mg/dL	Drug abuse	
ORGAN DYSFUNCTION		Malnutrition	
Cardiovascular	Binary (Presence)	ROUTINE CHARTED DATA	
Neurologic		Temperature	°C
Hepatic		Heart Rate	bpm
Hematologic		Arterial Blood Pressure Systolic	mmHg
Renal		Arterial Blood Pressure Diastolic	mmHg
Mechanical Ventilation		Arterial Blood Pressure Mean	mmHg
DATA TAKEN AT THE TIME OF ICU ADMISSION		Urine Output	mL
Gender	F, M	Base Excess	mEq/L
Admission Type	M,SS,US	Glucose	mg/dL
Age	Years	Peripheral capillary oxygen saturation (SpO2)	%
Glasgow Coma Scale (GCS)	Integer 3-15	Partial pressure of oxygen in arterial blood (PaO2)	mmHg

The variables of cohort A were obtained from various clinical and administrative values and were selected based on bibliographic revision, the criterion of the expert intensivists and the availability in the MIMIC-III database:

- Laboratory measurements: platelet count, bilirubin, creatinine, fraction of inspired oxygen, partial arterial pressure oxygen and fraction of inspired oxygen ratio, white blood cell count, potassium, sodium, bicarbonate, lactate, arterial pH, hematocrit and hemoglobin.
- Routine charted data: temperature, heart rate, arterial blood systolic pressure, arterial blood diastolic pressure, arterial blood mean pressure, urine output, base excess, glucose, peripheral capillary oxygen saturation and partial pressure of oxygen in arterial blood.
- Data taken at the time of ICU admission: gender, admission type, age and minimum Glasgow coma scale. Fourth the following comorbidities were extracted: diabetes, immunosuppressive diseases, malignancy, hematologic malignancy, metastatic cancer, heart failure, pulmonary diseases,

vascular diseases, coronary diseases, obesity, alcohol abuse, collagen diseases, drug abuse and malnutrition.

- The specific acute organ dysfunctions presented in each admission: (cardiovascular, neurologic, hepatic, hematologic, renal, respiratory).

Cohort A was extracted from the mentioned 15.254 admissions; For this, we selected the admissions with a hospital stays longer than 24 hours, resulting in a dataset with 13.836 patients. Then, only the admissions that had at least 70% of the laboratory measurements and at least 70% of routine charted data listed before were included in the study cohort A, getting 5.650 admissions. Figure 5.5 presents the accrual of admissions included in the study cohort and Figure 5.6 shows the distribution of admissions according to the used sepsis criteria.

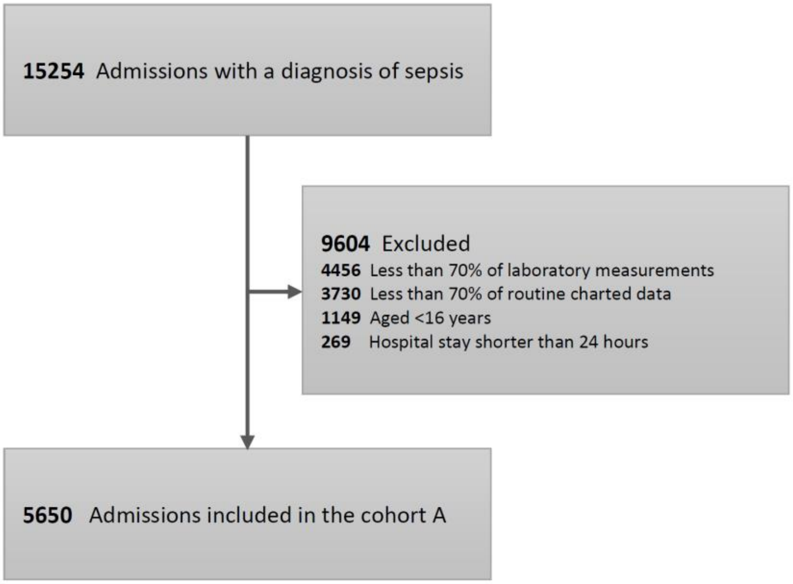

15254 Admissions with a diagnosis of sepsis
9604 Excluded
4456 Less than 70% of laboratory measurements
3730 Less than 70% of routine charted data
1149 Aged <16 years
269 Hospital stay shorter than 24 hours
5650 Admissions included in the cohort A

Figure 5.5. Accrual of admissions included in the study cohort A.

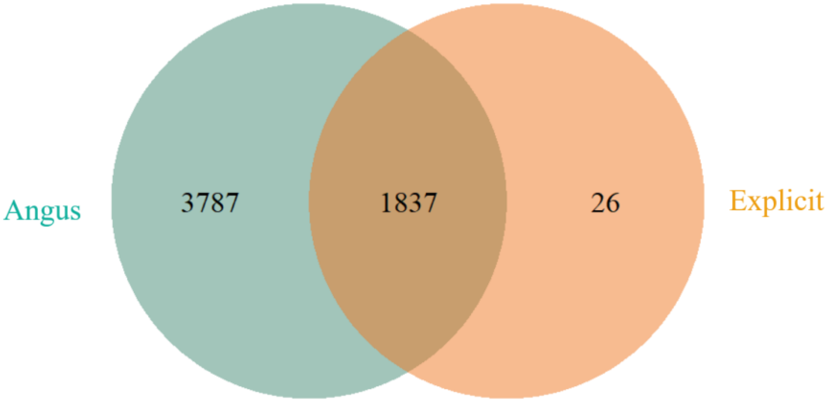


Figure 5.6. Venn Diagram of the study cohort A according to the used sepsis criteria.

Of all variables listed in Table 5.1, only four presented more than 5% of missing data being Bilirubin the most critical with 34% of absent values, followed by Fraction of Inspired O2 with 15%, Lactate with 13% and Base excess with 7%. The way in which the missing data are treated will be presented in Chapter 6.

To end this section it is worth noting that in the case of the cohort A, the majority of admissions were excluded due to the fact that at least 70% of routine charted data or laboratory measurements were not available; however, the studies conducted with this cohort (which will be presented in detail in later chapters) showed that not all included variables were relevant for the prediction of one-year mortality of ICU patients with sepsis. For this reason, the admissions of the cohort B were based on the variables that were relevant to the analyzes performed on the cohort A, and on those variables that most patients had.

5.3.2 Cohort B

From the 58977 MIMIC-III general population admissions, we selected all the ones that fulfill any of the four criteria presented in section 6.2. As result, 16.219 admissions were obtained; from these, we exclude the newborn patients (obtaining 16.080 admissions); finally, we only selected hospital admissions that where longer than one day, obtaining 15.751 admissions.

Since the objective of the studies carried out on the cohort B was to generate models based on the similarity between patients, we also included all the comorbidity categories present in the Elixhauser Comorbidity Index; a method of categorizing comorbidities of patients based on the ICD diagnosis codes. Each Elixhauser comorbidity category is dichotomous, which means it is either present or it is not. The Index can be used to predict hospital resource use and in-hospital mortality [87, 88]. According to the above, we extracted the clinical and administrative variables presented in Table 5.2.

After the variable extraction process, the admissions that did not have vital signs or laboratory measurements during the first 24 hours were also excluded, resulting in a study cohort of 15.476 admissions. Then, only the admissions that had at least 70% of the laboratory measurements and at least 70% of vital sings data listed before were included in the study cohort B, getting 15.082 admissions. Figure 5.7 presents the accrual of admissions included in the study cohort B and Figure 5.8 shows the distribution of admissions according to the used sepsis criteria.

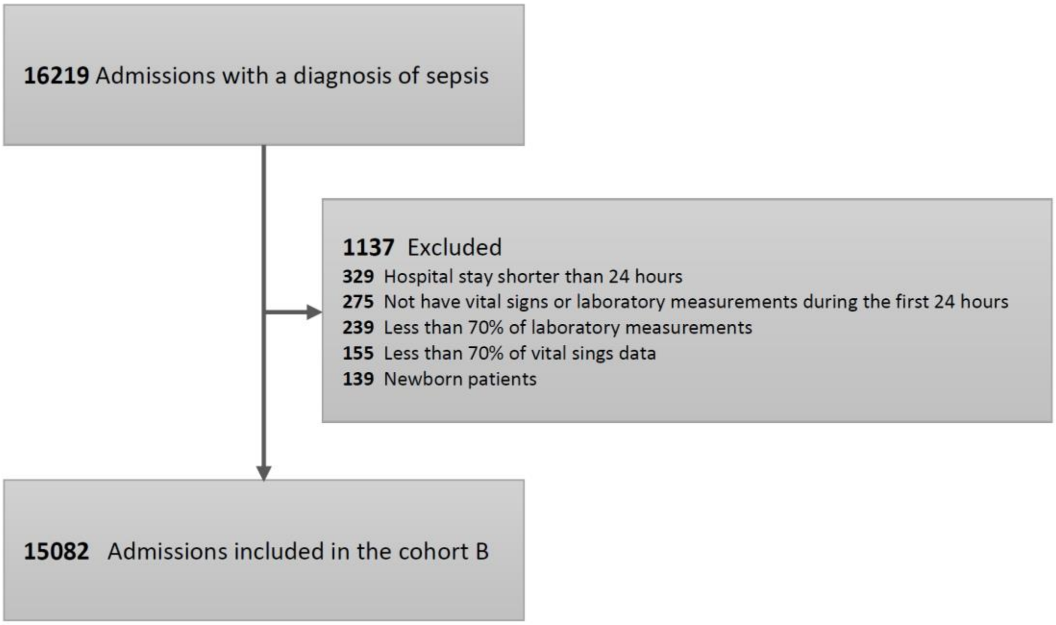

16219 Admissions with a diagnosis of sepsis
1137 Excluded
329 Hospital stay shorter than 24 hours
275 Not have vital signs or laboratory measurements during the first 24 hours
239 Less than 70% of laboratory measurements
155 Less than 70% of vital sings data
139 Newborn patients
15082 Admissions included in the cohort B

Figure 5.7. Accrual of admissions included in the study cohort B.

Table 5.2. Extracted predictors for Cohort B

Parameter	Unit	Parameter	Unit
Comorbidities		Data taken at the time of ICU admission	
Congestive heart failure	Binary (Presence)	Admission type	M,SS,US
Cardiac arrhythmias		Gender	F, M
Valvular disease		Age	Years
Pulmonary circulation		GCS	3-15
Peripheral vascular		Laboratory measurements	
Hypertension		Arterial pH	pH
Paralysis		Anion gap	mEq/L
Other neurological		Bicarbonate	mEq/L
Chronic pulmonary		Bilirubin	mg/dL
Diabetes uncomplicated		Creatinine	mg/dL
Diabetes complicated		Chloride	MEq/L
Hypothyroidism		Hematocrit	%
Renal failure		Hemoglobin	mg/dL
Liver disease		Lactate	mg/dL
Peptic ulcer		Platelet Count	10^9/L
Aids		Potassium	mEq/L
Lymphoma		Partial thromboplastin time (PTT)	s
Metastatic cancer		International normalized ratio (INR)	Ratio
Solid tumor		Prothrombin time (PT)	s
Rheumatoid arthritis		Sodium	mEq/L
Coagulopathy		Blood urea nitrogen (BUN)	mg/dL
Obesity		White Blood Cell (WBC) count	10^3/mm^3
Weight loss		Routine charted data	
Fluid electrolyte		Urine output	mL
Blood loss anemia		Heart rate	bpm
Deficiency anemias		Arterial Blood Pressure Systolic	mmHg
Alcohol abuse		Arterial Blood Pressure Diastolic	mmHg
Drug abuse		Arterial Blood Pressure Mean	mmHg
Psychoses		Respiratory rate	bpm
Depression		Temperature	°C
Treatments		Peripheral capillary oxygen saturation (SpO2)	%
Renal replacement therapy	Binary (Presence)	Glucose	mg/dL
Mechanical ventilation			

Of all variables extracted for cohort B, only three presented more than 5% of missing data being Bilirubin the most critical with 35% of absent values, followed by lactate with 21%, pH with 21% and ptt with 7%, inr-pt with 7%.

5.3.3 Data Preparation

Similar to the SAPS score, the routine charted data and laboratory measurements for both cohorts were extracted during the first 24 hours of each ICU stay; the other predictor variables represent single values throughout the entire duration of a patient ICU stay. Since the variables are not measured with the same frequency (Figure 5.9 depicts the time window for two of the variables as an example), we calculated statistical indices that allowed their description.

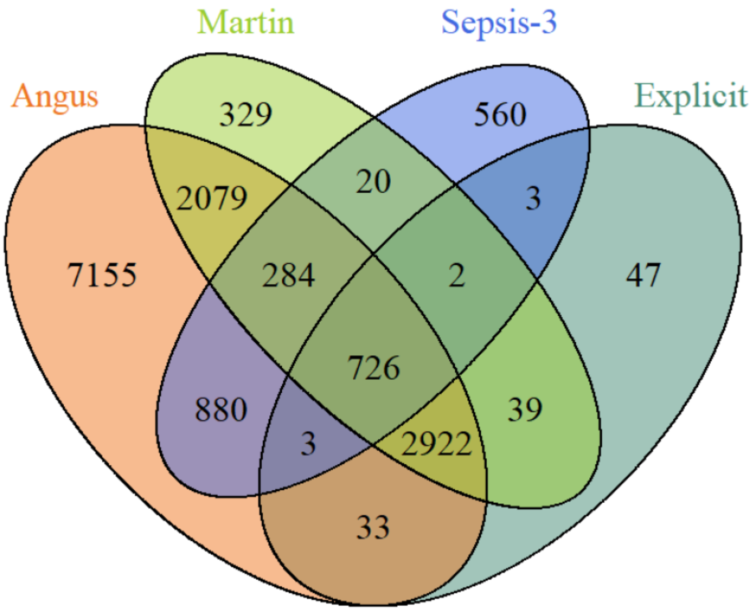


Figure 5.8. Venn Diagram of the study cohort B according to the used sepsis criteria.

For the cohort A we extract mean, maximum, minimum, variance and range for all the laboratory and routine charted variables. For the cohort B we only extracted maximum, minimum and mean values for vital signs, and maximum and minimum values for laboratory measurements. The reduction in the number of statistical indicators of cohort B was made because the analyzes performed on the cohort A indicated that the inclusion of the range and the variance did not significantly improve the performance of the mortality prediction models, and the clinical operationalization within the average intensive care units indicates that these values (and more complex ones such as kurtosis or skewness) could not be calculated.

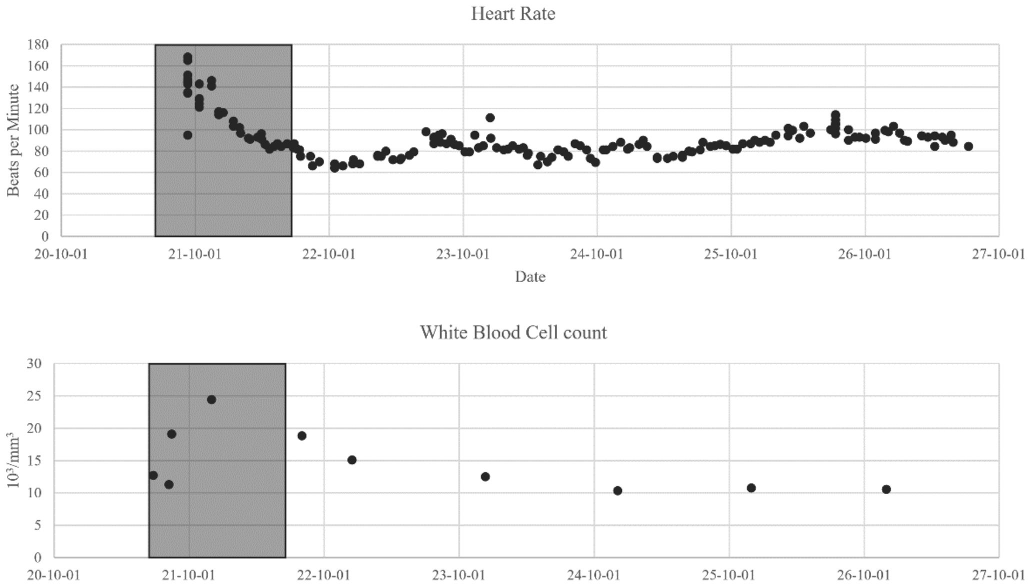


Figure 5.9. Example variables. Grey box represents the 24-hour window in which the data are extracted and evaluated.

5.4 Patient characteristics in the study cohorts

Table 5.3 . provides for cohort A a breakdown of the adult population by care unit. Cohort A contains data associated with 5.650 distinct hospital admissions for patients aged 16 or above who were given a severe sepsis or septic shock explicit diagnostic, or retrospectively identified as septic with the Angus criteria. The median age is 67.54 years, 54.58% of the patients are male, in-hospital mortality is 22.6% and one-year mortality is 43.36%. The median length of an ICU stay is 5.9 days and the median length of a hospital stay is 11.88 days

Table 5.3. Details of the cohort A patient population by first critical care unit on hospital admission. CCU is Coronary Care Unit; CSRU is Cardiac Surgery Recovery Unit; MICU is Medical Intensive Care Unit; SICU is Surgical Intensive Care Unit; TSICU is Trauma Surgical Intensive Care Unit.

	MICU	SICU	CCU	CSRU	TSICU	Total
Hospital Admissions	3138 (55.54%)	765 (13.54%)	735 (13.01%)	404 (7.15%)	608 (10.76%)	5650 (100%)
Different ICU stays	3402 (53.64%)	934 (14.73%)	828 (13.06%)	483 (7.62%)	695 (10.96%)	6342 (100%)
Age, median years	67.5	64.72	71.75	70.36	61.63	67.54
Gender(M)	1642 (52.32%)	393 (51.37%)	406 (55.23%)	248 (61.38%)	395 (64.96%)	3084 (54.58%)
ICU length of stay, median days	5.06	6.68	5.81	8	7.88	5.9
Hospital length of stay, median days	10.29	14.99	10.63	15.88	17.13	11.88
Hospital mortality	757 (24.12%)	165 (21.56%)	168 (22.85%)	76 (18.81%)	111 (18.25%)	1277 (22.6%)
One-year mortality	1459 (46.49%)	301 (39.34%)	346 (47.07%)	161 (39.85%)	183 (30.09%)	2450 (43.36%)

Table 5.4 provides for cohort B a breakdown of the adult population by care unit. Cohort B contains data associated with 1.582 distinct hospital admissions for patients aged 16 or above who were given a severe sepsis or septic shock explicit diagnostic, or retrospectively identified as septic with the Angus, Martin or Sepsis-3 criteria. The median age is 68.47 years, 53.87% of the patients are male, in-hospital mortality is 19.63% and one-year mortality is 42.92%. The median length of an ICU stay is 4 days and the median length of a hospital stay is 11.51 days.

Table 5.4. Details of the cohort B patient population by first critical care unit on hospital admission. CCU is Coronary Care Unit; CSRU is Cardiac Surgery Recovery Unit; MICU is Medical Intensive Care Unit; SICU is Surgical Intensive Care Unit; TSICU is Trauma Surgical Intensive Care Unit.

	MICU	SICU	CCU	CSRU	TSICU	Total
Hospital Admissions	8303 (55.05%)	2197 (14.57%)	1932 (12.81%)	1325 (8.79%)	1325 (8.79%)	15082 (100%)
Different ICU stays	9343 (53.79%)	2645 (15.23%)	2231 (12.84%)	1552 (8.93%)	1599 (9.21%)	17370 (100%)
Age, median years	67.78	65.93	72.84	71.84	65.85	68.47
Gender(M)	4328 (52.13%)	1183 (53.84%)	1039 (53.77%)	782 (59.02%)	794 (59.92%)	8126 (53.87%)
ICU length of stay, median days	3.28	5.24	4.2	5.8	6.02	4
Hospital length of stay, median days	9.68	15.13	10.65	15.16	15.65	11.51
Hospital mortality	1748 (21.05%)	413 (18.79%)	385 (19.93%)	195 (14.72%)	220 (16.60%)	2961 (19.63%)
One-year mortality	3847 (46.33%)	850 (38.69%)	887 (45.91%)	448 (33.81%)	442 (33.35%)	6474 (42.92%)

5.5 Conclusions

In chapter 4, we present the MIMIC-III clinical database. It contains data associated with 49,785 different hospital admissions for patients older than 16 years. The median age for adult general population is 65.8 years, in-hospital mortality rate is 11.5%, the median length of stay in the ICU is 2.1 days, and the median length of hospital stay is 6.9 days.

When comparing the general population with the population of selected sepsis cohorts, it can be observed that patients with sepsis have a hospital mortality, a median ICU length of stay and a median hospital length of stay that are close to twice the values for the general population.

The foregoing is true for both cohort A and cohort B; Moreover, it can be observed that although the sepsis identification methodologies yielded a different set of admissions, the inclusion of Martin and Spsis3 criteria do not significantly change the description in the cohorts.

In chapter 2, we presented some studies that indicates that patients with sepsis show ongoing mortality beyond the hospital discharge [56, 89–91]; which is ratified in the selected study cohorts, since one-year mortality rate is twice the in-hospital mortality rate.

From the accrual of admissions presented for cohorts A and B, it can be seen that the percentage of excluded admissions in cohort A is much higher than the percentage of excluded admissions the other one; the main reason is the condition that the admissions must have at least 70% of the data (both laboratory and routine charted). According to the above, the great difference between the number of admissions excluded is due to the fact that in the first cohort we sought to analyze variables selected by two criteria: the bibliographical review and the experts' opinions and in the second cohort we focused on variables that were relevant for 1-year mortality prediction in patients with sepsis within the ICU (see chapter 6) or that were routinely measured.

PART 3: CUSTOMIZED MODELS

This part presents our first approaches for the development of one-year mortality prediction models for sepsis patients within the intensive care unit (ICU). The works presented in this part lead to the generation of customized models, since they exclusively use the data of our study cohorts, in addition we generate obtained adjusted models based on traditional severity of disease scoring systems to benchmark the performance of our models. Chapter 6 present the development of a Stochastic Gradient Boosting model, and the obtaining of a set of variables that are truly related to the long-term mortality of sepsis patients within the ICU. Chapter 7 presents the development of two customized scores that indicate a patient's one-year mortality risk.

CHAPTER 6. MODEL FOR ONE-YEAR MORTALITY PREDICTION IN PATIENTS ADMITTED TO AN INTENSIVE CARE UNIT WITH DIAGNOSIS OF SEPSIS

6.1 Introduction

In previous chapters we showed that that sepsis is a life-threatening organ dysfunction due to a dysregulated host response to infection and it is an important public health problem, which generates high costs for the health system and carries a high morbidity and mortality; moreover; sepsis survivors had higher risks of all-cause mortality at 1 year after discharge compared to the general population. In this chapter we present the development of a model that goes beyond the prediction of in-hospital mortality and alert those patients who may have a poor prognosis one-year after being discharged from the hospital.

The model for the prediction of one-year mortality of sepsis diagnosed patients within the ICU was developed using the admissions of the study cohort A and was based on an advanced ensemble supervised learning method denoted as Stochastic Gradient Boosting (SGB). SGB combines boosting with bootstrap averaging and has built-in feature selection since it reports the relative variable importance. In addition to this, we also used selected relevant predictors by using a method that performs both variable selection and regularization called Least Absolute Shrinkage and Selection Operator (LASSO).

Thereby, we developed and evaluated five SGB models: one with all the predictors available in the study cohort A, two of them with the predictors selected with each of the methods (SGB and LASSO), another one with the union of the selected predictors, and the final with the intersection of the selected predictors.

All five developed models outperformed commonly used severity-of-disease scoring systems (SAPSII, SOFA and OASIS). As comparison measurements between the developed model and the traditional systems we used the accuracy and the AUROC; the Hosmer-Lemeshow goodness of fit test was used on the model to verify its ability to provide a risk estimate that corresponds to the observed mortality (Calibration). The calibration of our models were adequate since the p-value for the Hosmer-Lemeshow goodness of-fit were considerably greater than 0.05. This model would help identify those patients at greatest risk, and will be the first step to detect signs of alarm from a worse long-term outcome.

6.2 Methodology

In this approach we used the 5.650 admissions of the study cohort A. According to what was presented in the previous chapter, the analyzed variables were extracted during the first 24 hours of each hospital stay; and, since the variables are not measured with the same frequency, we calculated statistical indices that allowed their description: mean, maximum, minimum, variance and range. As result, data of the cohort A, were converted into 132 predictors as follow:

- 110 are the statistical descriptions of the laboratory measurements and the routine charted data,
- 17 are the presences of comorbidities and organ dysfunctions,
- 2 are the numerical values for age and Glasgow Coma Score (GCS).
- 3 corresponded to the gender and admission type categorical data, since each of these variables were binarized using one hot encoding.

In order to explain the data extraction process, we present examples of how variables from the five categories listed above were obtained, based on the actual data of a single particular patient with a sepsis related admission. In Table 6.1 we present some information that can be obtained directly from the MIMIC-III database. From this information we obtain the gender (assigning a 1 if the patient is "Male" and a zero if the patient is "Female"), the admission type (which is one hot encoded to two variables "Emergency" and "Elective"), the admission age (subtracting the date of birth from the admission time) and the one-year outcome (subtracting the death time from the discharge time).

Table 6.1. Admission information for the example patient.

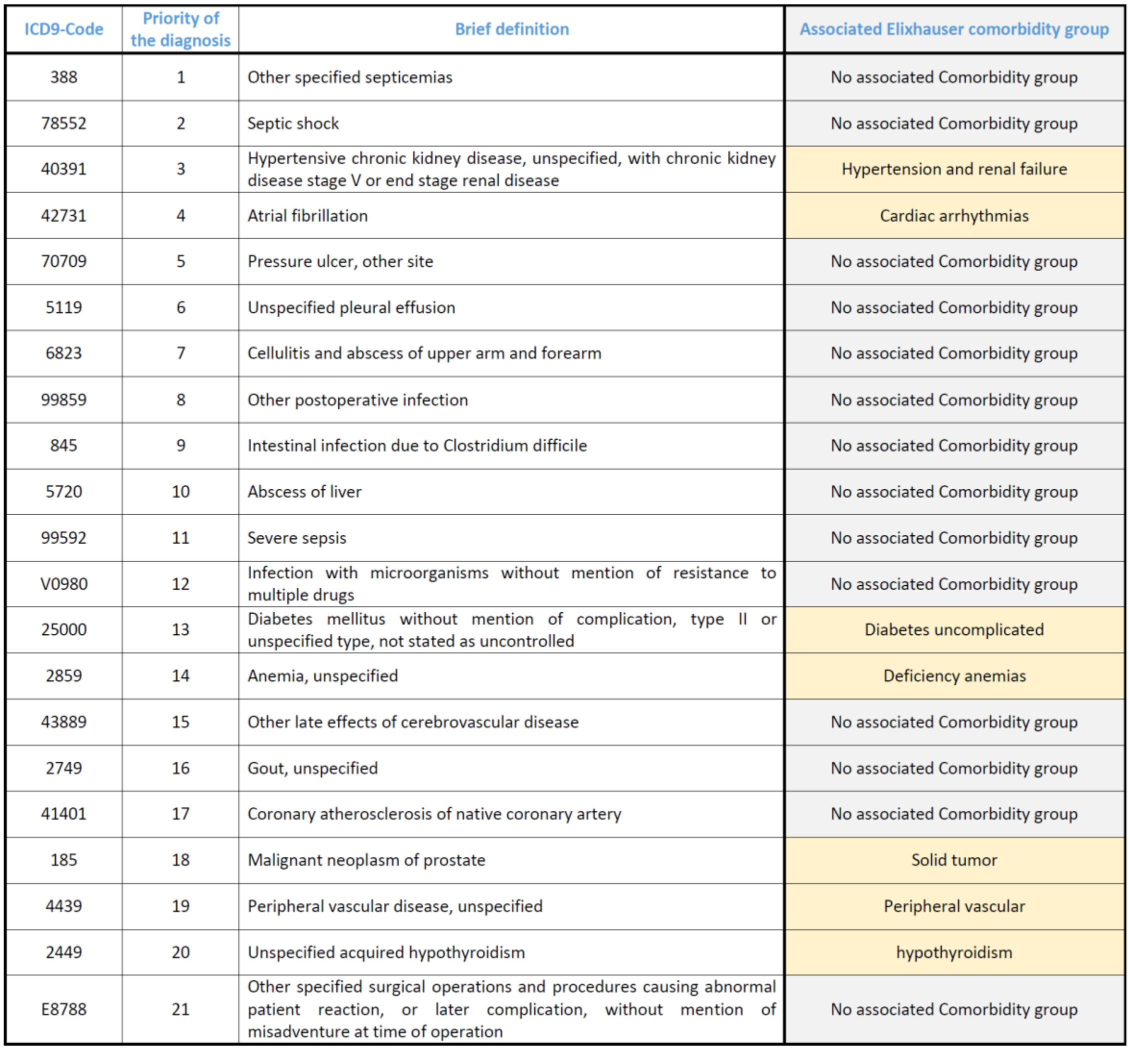
Admission time	Discharge time	Death time	Admission type	Diagnosis	Gender	Date of birth
30-01-35 20:50	08-02-35 2:08	08-02-35 2:08	EMERGENCY	SEPSIS	M	04-04-47 0:00

For the comorbidities we extract all the diagnoses that were made during the sepsis related admission; Table 6.2 presents the International Classification of Diseases Version 9 (ICD-9) codes that were registered for the sepsis related admission for our example patient; the three first columns can be obtained from directly from MIMIC-III, the fourth column is the assignation to an Elixhauser comorbidity group, not all the ICD-9 codes are related with a comorbidity group, and some of them are related with more than one group.

For the calculation of the minimum Glasgow Coma Scale (GCS), data associated with the three components (Eye opening response, verbal response and motor response) of the score for the first 24 hours of admission can be extracted directly from MIMIC-III, then, we add the numerical values of the three components and obtained the minimum one. Table 6.3 presents the information related to the GCS, the first three columns can be obtained directly from MIMIC-III, the fourth column is the associated behavior. It is important to note that MIMI-III provides the numeric value for the behavior and its interpretation; in orange we highlighted the data associated with the worst GCS that would be 13.

For each of the continuous numerical data of the laboratory measurements and the routine charted data we obtained all the records that were made to a patient within 24 hours after admission, and the we calculated the maximum, minimum, mean, variance and range. In Table 6.4 we present the values registered for the heart rate values for the example patients, in orange we highlighted the maximum a minimum value that were 83 and 66; the mean was 14.9, the range was 17 and the variance was 3.8.

Table 6.2. Comorbidity information for the example patient, it can be observed that in this admission eight comorbidities were registered.



ICD9-Code	Priority of the diagnosis	Brief definition	Associated Elixhauser comorbidity group
388	1	Other specified septicemias	No associated Comorbidity group
78552	2	Septic shock	No associated Comorbidity group
40391	3	Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage V or end stage renal disease	Hypertension and renal failure
42731	4	Atrial fibrillation	Cardiac arrhythmias
70709	5	Pressure ulcer, other site	No associated Comorbidity group
5119	6	Unspecified pleural effusion	No associated Comorbidity group
6823	7	Cellulitis and abscess of upper arm and forearm	No associated Comorbidity group
99859	8	Other postoperative infection	No associated Comorbidity group
845	9	Intestinal infection due to Clostridium difficile	No associated Comorbidity group
5720	10	Abscess of liver	No associated Comorbidity group
99592	11	Severe sepsis	No associated Comorbidity group
V0980	12	Infection with microorganisms without mention of resistance to multiple drugs	No associated Comorbidity group
25000	13	Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled	Diabetes uncomplicated
2859	14	Anemia, unspecified	Deficiency anemias
43889	15	Other late effects of cerebrovascular disease	No associated Comorbidity group
2749	16	Gout, unspecified	No associated Comorbidity group
41401	17	Coronary atherosclerosis of native coronary artery	No associated Comorbidity group
185	18	Malignant neoplasm of prostate	Solid tumor
4439	19	Peripheral vascular disease, unspecified	Peripheral vascular
2449	20	Unspecified acquired hypothyroidism	hypothyroidism
E8788	21	Other specified surgical operations and procedures causing abnormal patient reaction, or later complication, without mention of misadventure at time of operation	No associated Comorbidity group

We used two techniques in order to select the most important predictors for the one-year mortality prediction model: Least Absolute Shrinkage and Selection Operator (LASSO) and the Stochastic Gradient Boosting (SGB) variable importance. Since LASSO is based on maximum likelihood logistic regression it is susceptible to missing values, for this reason we used mean imputation. As an ensemble method based on decision tree aggregation, SGB can be fitted with even with the presence of missing values, therefore, there was no need for data imputation with this methodology. SGB variable importance is a procedure that indicates the contributions of each of the predictors to the model; therefore, it is possible to choose the most relevant predictors that represent the majority of the performance of the model.

After the feature selection process with, we developed five SGB models, two of them with the predictors selected with each of the methods, the third with the intersection of the predictors, the fourth with the union of the selected predictors with both methodologies and the last with all 132 predictors. The Hosmer–Lemeshow test assess whether or not the observed event rates match predicted event rates in subgroups of increasing probability of the one-year mortality.

Table 6.3. Glasgow Coma Scale related information for the example patient.

Chart time	Value	Value number	Behavior
30-01-35 22:00	4 Spontaneously	4	Eye opening response
30-01-35 22:00	6 Obeys Commands	6	Motor response
30-01-35 22:00	4 Confused	4	Verbal response
31-01-35 06:00	**3 To speech**	**3**	**Eye opening response**
31-01-35 06:00	**6 Obeys Commands**	**6**	**Motor response**
31-01-35 06:00	**4 Confused**	**4**	**Verbal response**
31-01-35 08:00	3 To speech	3	Eye opening response
31-01-35 08:00	6 Obeys Commands	6	Motor response
31-01-35 08:00	4 Confused	4	Verbal response
31-01-35 12:00	3 To speech	3	Eye opening response
31-01-35 12:00	6 Obeys Commands	6	Motor response
31-01-35 12:00	5 Oriented	5	Verbal response
31-01-35 16:00	3 To speech	3	Eye opening response
31-01-35 16:00	6 Obeys Commands	6	Motor response
31-01-35 16:00	5 Oriented	5	Verbal response
31-01-35 20:00	3 To speech	3	Eye opening response
31-01-35 20:00	6 Obeys Commands	6	Motor response
31-01-35 20:00	5 Oriented	5	Verbal response

Table 6.4. Heart rate measures for the first 24 hours after the admission of the example patient.

Chart time	Value	Units of measure	Chart time	Value	Units of measure
31-01-35 16:45	72	BPM	31-01-35 13:30	66	BPM
31-01-35 17:00	75	BPM	31-01-35 14:00	69	BPM
31-01-35 18:00	79	BPM	31-01-35 14:30	75	BPM
31-01-35 1:00	72	BPM	31-01-35 15:00	73	BPM
31-01-35 6:00	82	BPM	31-01-35 15:15	72	BPM
31-01-35 7:15	73	BPM	30-01-35 22:00	79	BPM
31-01-35 11:00	79	BPM	31-01-35 2:00	79	BPM
31-01-35 11:15	80	BPM	31-01-35 3:00	75	BPM
31-01-35 12:00	77	BPM	31-01-35 4:00	77	BPM
31-01-35 15:20	72	BPM	31-01-35 19:00	78	BPM
31-01-35 15:30	73	BPM	31-01-35 20:00	77	BPM
31-01-35 15:35	74	BPM	31-01-35 8:00	72	BPM
31-01-35 16:00	73	BPM	31-01-35 12:45	78	BPM
31-01-35 5:00	83	BPM	31-01-35 13:00	73	BPM
31-01-35 8:45	71	BPM	31-01-35 13:15	73	BPM
31-01-35 9:00	76	BPM	30-01-35 23:00	75	BPM
31-01-35 10:00	79	BPM	31-01-35 0:00	68	BPM
31-01-35 10:30	75	BPM			

6.2.1 SGB

SGB is a treebased ensemble-based algorithm, capable of manage qualitative and quantitative variables, and remain robust to missing data and outliers. SGB model has been successfully applied in various fields such as rockburst damage prediction, travel time prediction, land cover mapping and berries skin flavonoid contents [92–95], and even for prediction of mortality in head injury [96], where the Boosted Tree Classifier method achieved both the highest AUROC and accuracy rate.

Ensemble based algorithms consist of multiple base models, each one of those provides a different solution to the problem. The solutions of all the base models are finally combined (usually by weighted voting or averaging) into a single final model output, that is usually a more stable and accurate prediction [97]. The ensemble algorithms begin with a training dataset:

$$\{y_i, x_i\}_1^N$$

which Where y_i is the response variable and $x_i = \{x_1, x_2, \dots, x_n\}$ is a feature vector with n variables and the whole dataset is composed of N observations; and the goal is to find a function $F^*(x)$ that maps x to y and minimizes a loss function $\Psi\big(y, F(x)\big)$ [98].

Stochastic Gradient Boosting (SGB) is a type of ensemble algorithm based on Gradient Boosting [99], a function approximation method that estimates $F^{*}(x)$ using an additive expansion:

$$F(x) = \sum_{m=0}^{M} \beta_m h(x; a_m) \ (1)$$

Where $h(x; a)$ is a simple parameterized function of the input variables x, characterized by a set of parameters $a = \{a_1, a_2, \dots\}$. The individual terms of the function $h(x; a)$ differ in the values chosen for the parameters a_m. The expansion coefficient $\{\beta_m\}_0^M$ and the model parameters $\{a_m\}_0^M$ must be jointly fitted to the training data [98].

The algorithm stars with guess of F_0; after that the expansion coefficient, the parameters and the function are calculated iteratively for $m = 1,2, \cdots, M$ as following:

$$(\beta_m, a_m) = \arg\min_{\beta,a} \sum_{i=1}^{N} \Psi\big(y_i, F_{m-1}(x_i) + \beta h(x_i; a)\big) \ (2)$$

and

$$F_m(x) = F_{m-1}(x) + \beta_m h(x_i; a_m) \ (3)$$

F_m represents the function in the m^th^ iteration and β_m and a_m are the expansion coefficient and the parameters of the m^th^ simple model. In order to solve (2) for arbitrary differentiable loss function $\Psi\big(y, F(x)\big)$ a two-step procedure have been proposed [99]. For this, first, the function $h(x; a)$ is fitted by least-squares:

$$a_m = \arg\min_{a,\beta} \sum_{i=1}^{N} [\tilde{y}_{im} - \beta h(x_i; a)]^2 \ \ (4)$$

In (4) $\tilde{y}_{im}$ is:

$$\tilde{y}_{im} = -\left[\frac{\partial \Psi\big(y_i, F(x_i)\big)}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \ (5)$$

Then given $h(x; a)$, the optimal value of β_m is calculated:

$$\beta_m = \arg\min_{\beta} \sum_{i=1}^{N} \Psi\big(y_i, F_{m-1}(x_i) + \beta h(x_i; a_m)\big) \ (6)$$

After that, the updated function $F_m$is calculated:

$$F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m) \ \ (7)$$

Particularly SGB uses decision trees as base model [100]. Thus, $h(x; a)$ is a L-terminal node small regression tree (with 3 to 9 splits for this study); at each iteration, the regression tree divides the explanatory variables space into L-disjoint sub-regions $\{R_{lm}\}_{l=1}^{L}$ in each of which a constant response value $\bar{y}_{lm}$ is calculated, moreover, with regression trees (6) can be solved separately with in each sub-region R_{lm} so it solution reduces to:

$$\gamma_{lm} = \arg\min_{\gamma} \textstyle\sum_{x_i \in R_{lm}} \Psi(y_i, F_{m-1}(x_i) + \gamma) \ (8)$$

Where γ_{lm} is an estimate of the expansion coefficient in a particular sub-region. And the current approccimation of the function is updated in each corresponding region by:

$$F_m(x) = F_{m-1}(x) + \nu \cdot \gamma_{lm} 1(x \in R_{lm}) \ (9)$$

Where $\gamma_{lm} 1(x \in R_{lm})$ is the a constant response value in a particular sub-region and ν is a shrinkage parameter that controls the learning rate of the procedure. Theoretically ν must be $0 < \nu \leq 1$, but it has been, empirically, it was found that small values ($\nu \leq 0.1$) lead to much better generalization [99].

To improve the performance of the gradient boosting approach, SGB incorporates randomness into the function estimation procedure, so at each iteration a random permutation $\{\pi(i)\}_1^N$ is selected (without replacement and with size $\widetilde{N}$) This randomly selected subsample is then used, instead of the full sample, to fit the decision tree and compute the model update for the current iteration.

For this study the output variable y is binary, and the loss criterion is the deviance [98]

$$\Psi(y, \hat{F}) = 2\log(1 + \exp(-2y\hat{F})) \ (10)$$

The SGB algorithm involves a parameter-tuning process that maximizes predictive accuracy, the three parameters are: M is the total number of boosting iterations (n.trees), υ is the learning rate (shrinkage coefficient) and L is the number of splits performed on each tree [93]. To determine the optimal combination of the mentioned parameters, 10 fold cross-validation procedure was applied for each parameter configuration. In this procedure, the elements of the train subset were randomly divided into 10 groups, nine of these groups were selected for fitting a model and the other one was used for testing it. The process was repeated ten times, so each group was used for testing and training. By averaging the results produced in each iteration, an overall quality estimate was obtained. Finally, the combination of parameters that minimizes the prediction error averaged across all 10 folds was selected as the final model.

The validation subset was never used in the development of the SGB model, but it was used to evaluate the performance of the final model. Variable importance was calculated using the improvement based on the splitting criteria for each predictor, which are aggregated and averaged across the entire boosting ensemble [92, 93, 100, 101].

6.2.2 LASSO

LASSO [102], is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model. The method for two-class classification seeks $p(X)$ the probability of class membership and is based on a hypothesis function that lies between 0 and 1. For logistic regression [103]:

$$p(X) = \frac{exp^{(\beta_0 + \beta_1^T)}}{1 + exp^{(\beta_0 + \beta_1^T)}} \qquad (11)$$

Which is equivalent to:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1^T \quad (12)$$

Where β_1 is a vector, with as many components as there are predictors, and the objective is to find the values of β_1 that results in a *p(X)* that most accurately classifies all the observed data points. Logistic regression models can be fitted by maximum likelihood. The log-likelihood can be written:

$$l = \sum_{i=1}^{N} \left\{ y_i(\beta_0 + \beta_1^T x_i) - \log\left(1 + e^{(\beta_0 + \beta_1^T x_i)}\right) \right\} \quad (13)$$

LASSO regularization works by adding a penalty term to the log likelihood function, thus the quantity to be minimized is:

$$l + \lambda \sum_{j=1}^{p} |\beta_{1j}| \quad (14)$$

λ is a complexity parameter that controls the amount of shrinkage, so, the larger the value of λ, the greater the amount of shrinkage. λ is selected using cross validation in a way that the resulting model minimizes the sample error. The effect of the LASSO penalty term is to set some of the models coefficients exactly to zero, and thus allowing the variable selection.

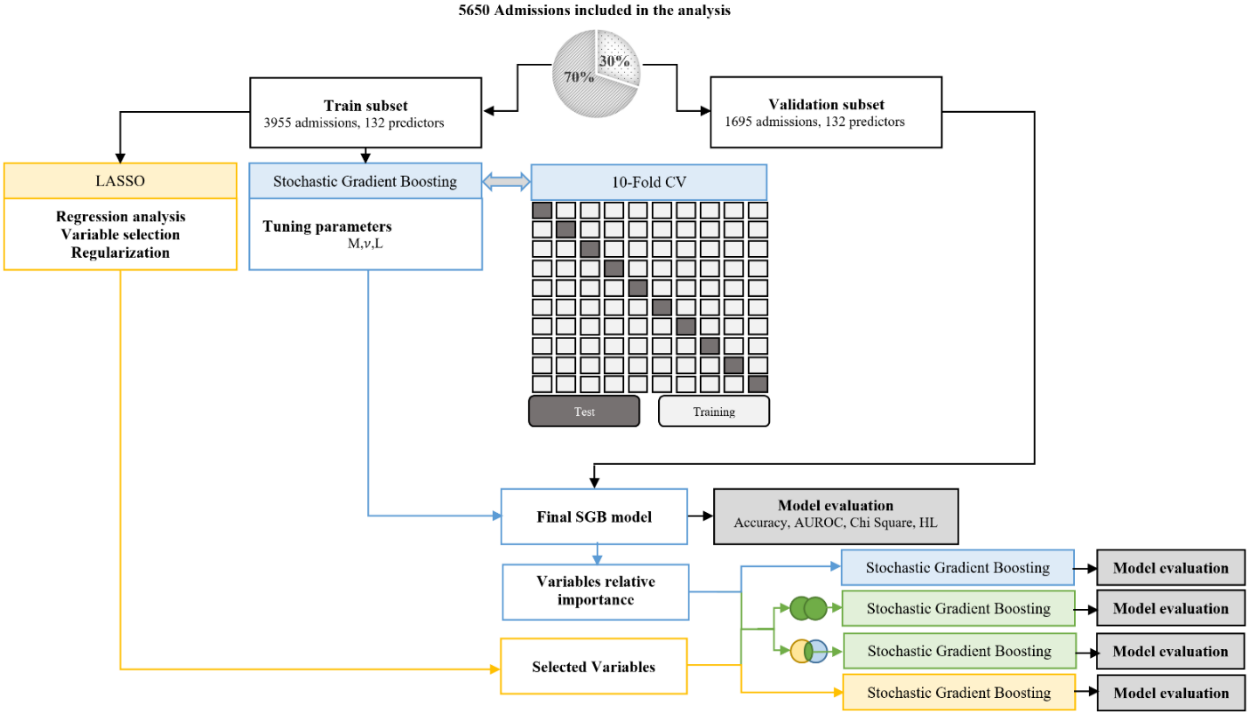

5650 Admissions included in the analysis
30%
70%
Train subset
3955 admissions, 132 predictors
Validation subset
1695 admissions, 132 predictors
LASSO
Regression analysis
Variable selection
Regularization
Stochastic Gradient Boosting
Tuning parameters
M,ν,L
10-Fold CV
Test
Training
Final SGB model
Model evaluation
Accuracy, AUROC, Chi Square, HL
Variables relative importance
Selected Variables
Stochastic Gradient Boosting
Model evaluation
Stochastic Gradient Boosting
Model evaluation
Stochastic Gradient Boosting
Model evaluation
Stochastic Gradient Boosting
Model evaluation

Figure 6.1 Model development

6.2.3 Performance

In order to develop the models, we used the following methodology. First we divided the study cohort A into 70% for training and 30% for testing. With the training subset we extracted two subsets of features using the SGB and LASSO methodologies. The parameters of each of those methodologies were tuned using ten-fold cross validation. Finally, we developed five SGB models with different set of predictors: the 132 predictors, the SGB variable importance selected predictors, the LASSO selected predictors, the union and interception of the selected predictors with both methodologies. The whole process was repeated 50 times in order to obtain a confidence interval. Figure 6.1 illustrates the methodology.

To evaluate the performance of the five SGB models with the different sets of predictors we use the area under the Receiver Operating Characteristic curve (AUROC) that is a common indicator of the goodness of a predictor in a binary classification task, and the accuracy, which is defined as the fraction of correctly predicted records over the number of records. These measures of the predictive power of SGB models were compared with three severity-of-disease classification systems: SOFA [104, 105], SAPS2 [5] and OASIS [106].

Additionally, the developed models calibration was assessed using Hosmer-Lemeshow test. The Hosmer–Lemeshow test is a commonly used procedure for assessing goodness of fit in binary classification problems; It is widely used for the evaluation of risk-scoring models in medicine that are developed using a wide range of sample sizes [107]. The Hosmer–Lemeshow test seeks to prove that a model fits the data, and it is a chi-square test conducted by sorting the n observations in the data set by estimated probability of success, dividing the sorted set into g groups and assessing the Hosmer–Lemeshow C statistic:

$$\hat{C}_g = \sum_{i=1}^{g}\left[\frac{\left(O_{s,i}-E_{s,i}\right)^2}{E_{s,i}} + \frac{\left(O_{f,i}-E_{f,i}\right)^2}{E_{f,i}}\right] \quad (14)$$

Where $O_{s,i}$ and $O_{f,i}$ are the observed number of successes and failures; $E_{s,i}$ and $E_{f,i}$ the predicted successes and failures in the ith group; $\hat{C}_g$ follows a Chi-square distribution with $g-2$ degrees of freedom, therefore the p-value for the test is:

$$p = \int_{\hat{C}_g}^{\infty} \chi^2_{g-2}(x)dx \quad (15)$$

The number of groups g is defined by the user and there is an established dependence of the probability of correctly rejecting a poorly fitting model (using the Hosmer-Lemeshow test) with the sample size and the number of groups. By considering this dependence, Paul et al. made some recommendations to select the number of groups. Specifically for samples sizes between 1000 and 25000 g is given by:

$$g = max\left(10, min\left\{\frac{m}{2}, \frac{n-m}{2}, 2+8\left(\frac{n}{1000}\right)^2\right\}\right)$$

Where n is the number of observations, and m is the number of successes [107]. However, this recommendation is based on the simulation of six models, all of which are much simpler than the SGB models developed in this work.

Finally, since, calibration, and therefore the Hosmer-Lemeshow test, examines how well an observed number of events (deaths) compare to the number of events estimated by the model across probability groups, we compared, to give a friendly visual representation, observed versus model-based estimates of numbers of deaths graphically within deciles (g=10) of estimated probabilities of mortality.

6.3 Results

The SGB models were implemented with the caret [108] and gbm [109] R-packages, LASSO was implemented using glmnet [110] R-package; both in R software. 10-fold CV procedure was used to tune the parameters. According to what was previously mentioned, for SGB, three parameters were tuned: M, υ and L. To determine the combination of parameters that present better performance (greater AUROC) during the CV process, a set of SGB models were tested using different values for M (50, 100, 150, 200, 250, …, 1900, 1950, 2000), L (3, 5, 7, 9) and υ (0.001, 0.01, 0.1).

The first developed model was fitted using all the predictors. Figure 6.2 presents the tuning procedure for this model parameters that was done using the results of the performance of cross-validation; Each of the four plots in the figure represents the maximum number of splits performed on each tree (iteration depth, parameter L); The colored line in each of the four plots represent different shrinkage coefficient (learning rate, parameter υ); The x axis of each plot represent the boosting iteration (parameter M). Each data point in the figure represents one classifier. The optimal classifier is constructed with 950 trees, an iteration depth of 9 and a learning rate of 0.01; the AUROC obtained on the 3955 admissions of the training subset was 0.7715(95% Confidence Interval: 0.762 - 0.786) and the AUROC obtained on the validation subset was 0.805 (95% Confidence Interval: 0.785 - 0.826).

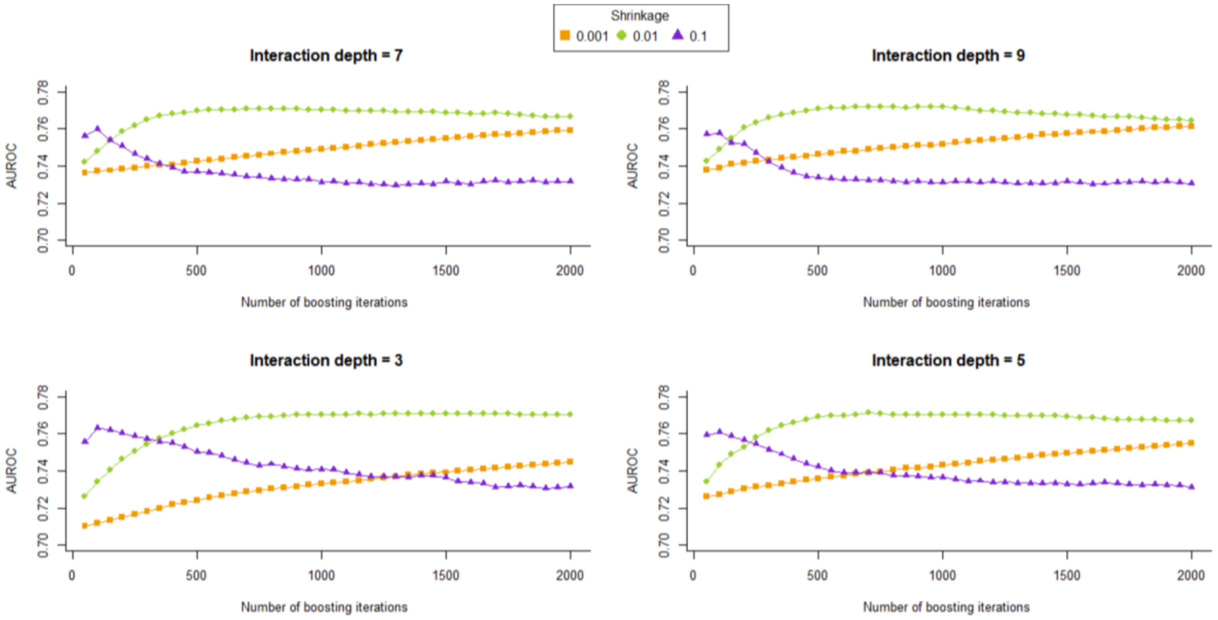


Figure 6.2 SGB model tuning parameters and AUROC. Each data point in the figure represents one classifier. For example, in the lower-left plot the purple data point at (400,0.746) indicates a model built with 400 trees, a tree depth is 3 and a learning rate of 0.1; this particular classifier gives an AUC value of 0.746 in the 10-fold cross-validation on the training subset.

A second SGB model that only used a subset variables selected using the relative variable importance method built in within the SGB. The general effect on the model of each predictor was calculated using the relative importance, this measures are based on the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees[100, 111]. The most influential predictors were selected by developing a model with the predictors that had the greatest relative importance, and that in the end presented an AUROC like that of the complete model. 40 predictors were selected (that represent around 68% of the influence in the model), the most important one, as expected was admission age with a relative importance around 10%; the second most relevant predictor was total urine output during the first day of admission, although urine output is a commonly used indicator for renal disease, it is interesting how much it affects the log-term mortality; the relative importance of the following predictors are below 4%. The AUROC obtained on the 3955 admissions of the training subset was 0.7713 (95% Confidence Interval: 0.766 - 0.7763) and the AUROC obtained on the validation subset was 0.803 (95% Confidence Interval: 0.783 - 0.824).

Figure 6.3 shows the relative importance of the selected predictors with the SGB methodology for the model developed with all the variables, and the model that only uses the selected ones; since the relative importance of the variables are scaled so that the sum adds to 100 [100], it is clear that, in the model that only uses the 40 SGB selected variables the relative importance of each variable increase with respect to the model with all the variables; however, although there is a change in the relative positions of the predictors, no particularly sharp change is observed.

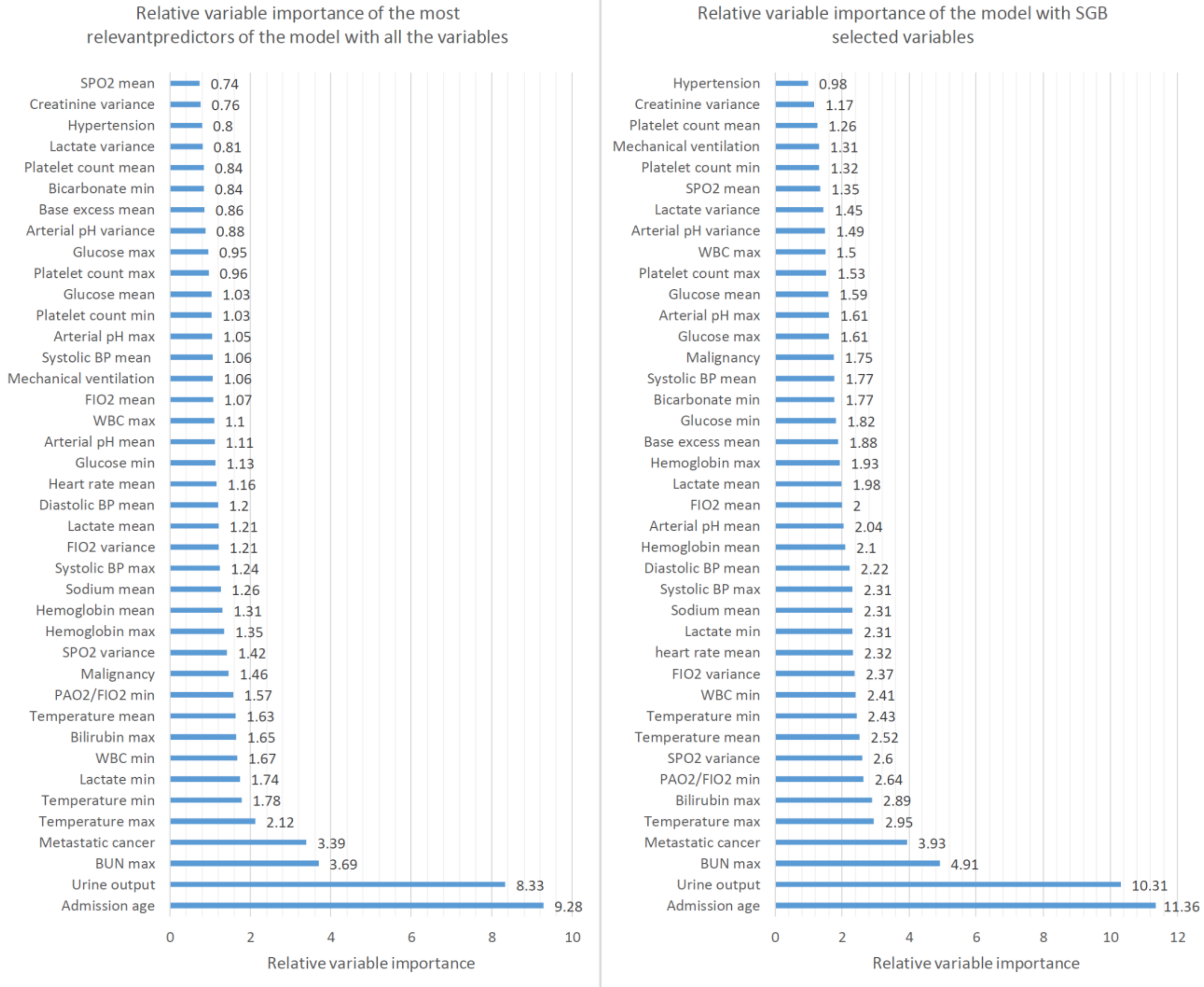


Figure 6.3 SGB relative importance of the predictors for 1-year mortality prediction models. In the left we presented the most relevant predictors of the complete model (that sums 67%); in the right we present the relative importance of the model developed only with selected predictors. Abbreviations: Bun: blood urea nitrogen; Max: maximum; WBC: white blood cell; Min: minimum; SpO2: peripheral capillary oxygen saturation; PaO2/FiO2: partial pressure arterial oxygen and fraction of inspired oxygen ratio; FiO2: fraction of inspired oxygen; Mechanical vent: mechanical ventilation; DABP: diastolic arterial blood pressure; BP: blood pressure.

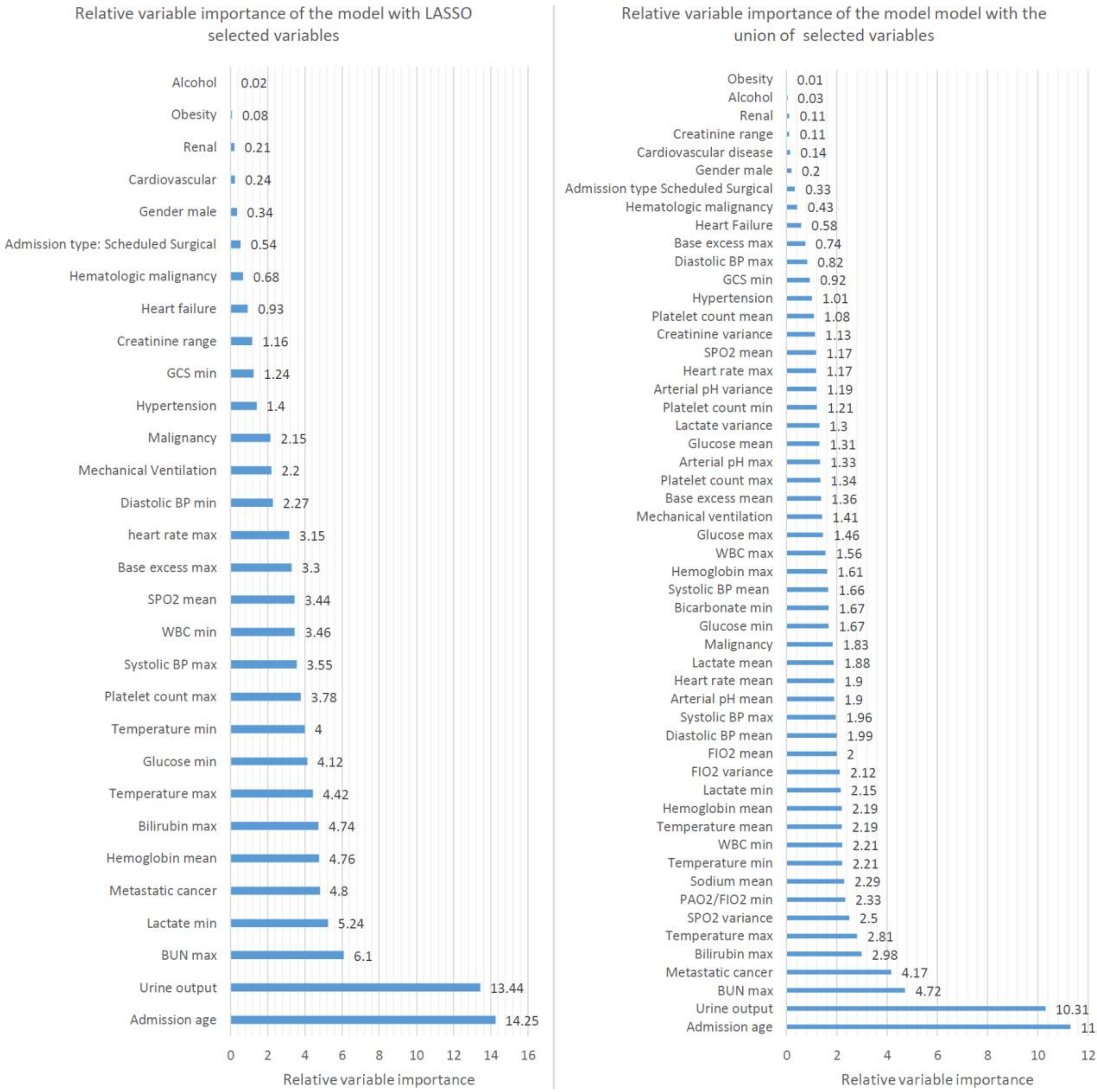


Figure 6.4 SGB relative importance of the predictors for 1-year mortality prediction models. In the left we presented the relative variable importance of the model developed with the LASSO selected predictors; in the right we present the relative variable importance of the model developed union of the selected predictors. Abbreviations: Bun: blood urea nitrogen; Max: maximum; WBC: white blood cell; Min: minimum; SpO2: peripheral capillary oxygen saturation; PaO2/FiO2: partial pressure arterial oxygen and fraction of inspired oxygen ratio; FiO2: fraction of inspired oxygen; Mechanical vent: mechanical ventilation; DABP: diastolic arterial blood pressure; BP: blood pressure; GCS: Glasgow coma scale.

The third SGB model was developed using only a subset of predictors obtained with the LASSO Methodology. Unlike the methodology based on SGB variable importance, LASSO sets some of the predictors coefficients to zero, whereby the algorithm returns a set of selected variables. In order to obtain a subset of values truly related to the one-year mortality, we performed 100 runs with different splits for the training and validation datasets (always 70% for training), each model was fitted with the random training subset, thus, on each run we obtained a slightly different subset of LASSO selected predictors. According to this, we developed the SGB model with the LASSO predictors that were selected in more than 80% of the runs. The AUROC obtained on the 3955 admissions of the training subset was 0.7714 (95% Confidence Interval: 0.765 - 0.777) and the AUROC obtained on the validation subset was 0.806 (95% Confidence Interval: 0.785 - 0.826).

The fourth SGB model was developed using the union of the variables selected with both SGB variable importance and LASSO methodologies. The AUROC obtained on the 3955 admissions of the training subset was 0.7715 (95% Confidence Interval: 0.766 - 0.779) and the AUROC obtained on the validation subset was 0.808 (95% Confidence Interval: 0.787 - 0.828).

Figure 6.4 shows the relative importance of the variables of SGB model developed with the LASSO selected variables, and the model with the union of the selected variables with both LASSO an SGB methodologies. It can be observed that, in contrast to the values reported in Figure 6.3, some predictors that were selected with LASSO (Alcohol, Obesity, Renal, Gender males among others) but not with SGB variable importance, do not significantly influence the one-year mortality prediction, being obesity and alcohol the most extreme cases.

According to the above, the final model was developed with the intersection of the selected predictors with both methodologies. Figure 6.5 shows the relative variable importance of the model with the intersection of selected variables, this intersection model leads to the development of a much simpler model that only has 17 predictors. The AUROC obtained on the 3955 admissions of the training subset was 0.754 (95% Confidence Interval: 0.746 - 0.759) and the AUROC obtained on the validation subset was 0.791 (95% Confidence Interval: 0.769 - 0.812).

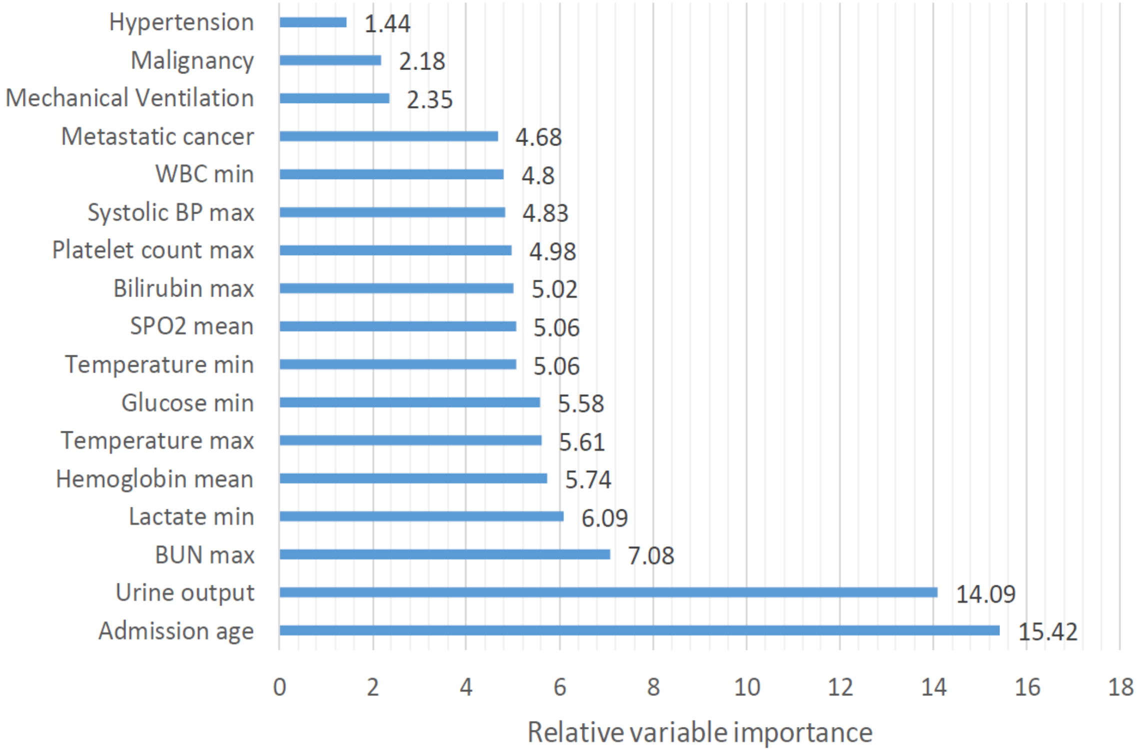


Figure 6.5 Relative variable importance of the model with the intersection of selected variables

Calibration of the five models were evaluated using Hosmer–Lemeshow Test (with g=25); and the parameters of all the models were fitted using a 10-fold cross validation process on a training subset; Table 6.5 summaries the number of variables, the final parameters values and the evaluation measures of the five models over the 1695 admissions of the validation subset.

*Table 6.5 Models performance measures. * the learning rate where maintained constant for all the models in υ=0.01*

Model	Predictors	Accuracy	AUROC	HL p-value	Parameters*
Complete	132	0.726	0.805	0.058	L=9 M=950
LASSO	30	0.736	0.806	0.353	L=9 M=700
SGB	40	0.736	0.803	0.081	L=9 M=800
Union	53	0.731	0.808	0.151	L=7 M=950
Intersection	17	0.721	0.791	0.212	L=9 M=800

It can be observed from Table 6.5, that SGB variable importance and LASSO methodologies allowed to develop SGB models that preserve the same performance as the model generated with all the predictors but with as subset of predictors, moreover, it can be seen that the 17 intersection variables are the ones that are truly related to the one-year mortality in sepsis patients since the model developed with them achieves a performance similar to other models. For the intersection model, observed versus predicted of numbers of deaths were compared graphically within deciles of increasing probability of the outcome. The graph presented in Figure 6.6 indicate that estimated and observed mortality pairs are similar and shows that the number of outcome events is indeed increasing along the probability deciles.

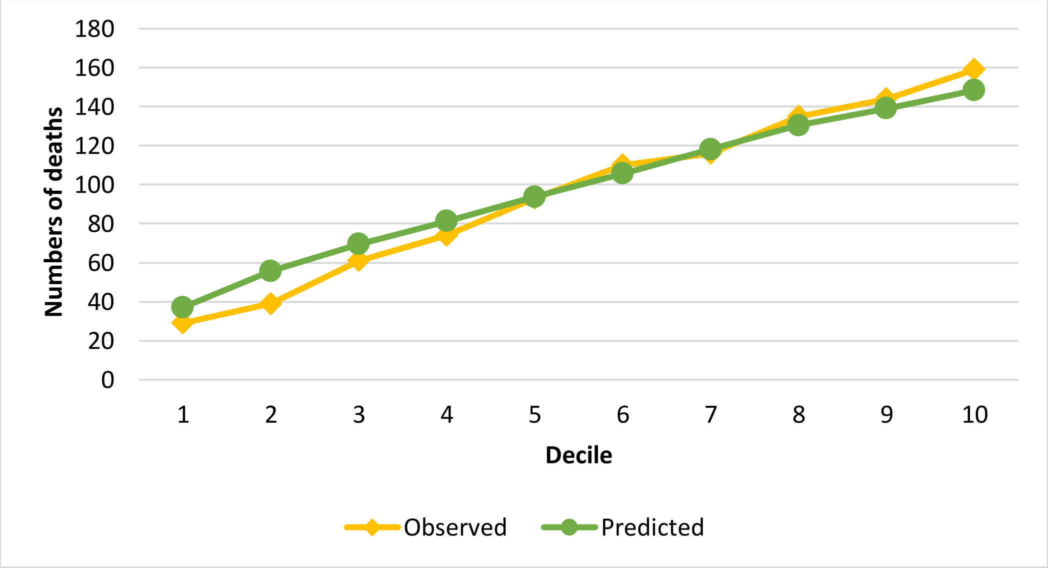
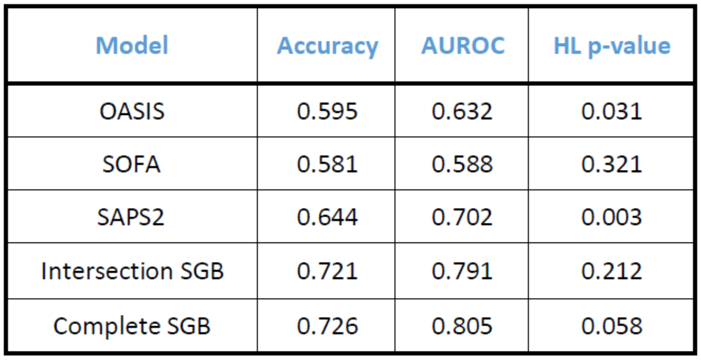


Figure 6.6 Comparison of observed versus predicted one-year mortality in the deciles of predicted mortality based on the SGB model with the intersection variables.

To benchmark the SGB model, we adjusted three severity-of-disease scoring systems (OASIS, SAPS II, and OASIS) and obtained their AUROC and Accuracy. The adjustment process was proposed by Le Gall et al. in [63], and consist to modify an existing severity of disease score by adapting them for use specifically among patients that shares a common characteristic. For instance, When the standard SAPS II model is applied to a patient, all of the variables are assigned points, and the resulting sum is the SAPS II score, which is then used as a variable in a logistic regression equation to generate the mortality probability. To adjust a model for patients with sepsis, it is necessary to develop new logistic regression equations using data only from the group of patients with sepsis. Such new model is a logistic regression that contains a single variable (The particular score that are being adjusted) plus the constant term. When this process is applied, the original score, which produced a probability of mortality for general medical population, would be mathematically translated into an adjusted probability of mortality based only on the experience of patients with sepsis. Le Gall et al proved that the adjusted models presented better calibration and discrimination than traditional scores [63]. The performance of the traditional severity of disease classification systems adjusted for the training subset and evaluated on the validation subset are presented in Table 6.6.

Table 6.6 AUROC, accuracy and calibration for the 1-year mortality on the validation subset for three adjusted severity of disease scoring systems and the proposed SGB models with all variables (132 predictors), and the variables from the intersection between the SGB variable importance selected variables and the LASSO selected variables (17 predictors).



Model	Accuracy	AUROC	HL p-value
OASIS	0.595	0.632	0.031
SOFA	0.581	0.588	0.321
SAPS2	0.644	0.702	0.003
Intersection SGB	0.721	0.791	0.212
Complete SGB	0.726	0.805	0.058

6.4 Conclusions

The presented models for the one-year mortality prediction of the patients that are admitted in a ICU with a sepsis diagnosis; shows that the use of ensemble based algorithms (SGB in this study) and the inclusion of predictors that are not usually taken into account in the traditional severity-of-disease classification systems (for example minimum lactate), outperform some traditional severity of disease scoring system for long-term mortality prediction task.

The fact that the developed SGB models presented a higher accuracy and AUROC over the validation subset ratifies that custom mortality prediction models for a specific disease presents a better performance that traditional severity of disease scoring systems, which could lead to better management of illness within the ICU. Although the SGB models presents a good interpretability, since they retrieve the relative importance of the predictors, it is clear that there is a complex interdependence among different physiological systems in response to sepsis, since the SGB models are composed of between 450 and 1150 trees; for this reason, it would be necessary to develop easy-to-use computer tools that allow these types of models to be implemented within the ICU.

The 17 predictors of the intersection model, allow to identified features that could become prognostic markers for the one-year mortality of the sepsis diagnosed patients within the ICU. As expected, older patients are at greater risk in consequence the most important parameter for the outcome is the age. Urine output is used as a marker of acute kidney injury, a disease that is associated with substantial in-hospital mortality [112].

Minimum lactate over the first 24 hours of the ICU admission is the seventh most important variable for the mortality prediction in this study; Lactate is currently used within the ICU as a diagnostic tool and as a prognostic marker, since the higher the value, the greater the risk of mortality. However, if the lactate of a patient does not reach below a threshold, it will also have a higher mortality risk. for this reason, the minimum lactate during the first 24 hours must also be analyzed in ICUs. Mean lactate is also considered an important predictor, which agrees with what is reported in the literature, since hyperlactatemia is related with a poor outcome in ICU [113]. An elevated Blood urea nitrogen (BUN) is associated with increased mortality in critically ill patients [114].

CHAPTER 7. SCORING SYSTEM FOR THE ONE-YEAR MORTALITY RISK OF SEPSIS PATIENTS IN INTENSIVE CARE UNITS AND STRATIFICATION OF PATIENTS IN RISK GROUPS

7.1 Introduction

In chapter 6 we showed that the use of machine learning techniques for the development of customized mortality prediction models, that use only the data of a population that shares a common characteristic, leads to better performance compared to the general population severity of disease classification systems. In addition, we found a subset of predictors that are truly related with the long-term mortality prediction in sepsis patients within the ICU; such subset of predictors was used to develop the study cohort B, which is composed of 15082 hospital admissions (representing the 93% of the total sepsis related admission of MIMIC-III), meaning that the conclusions driven from the study cohort B are more representative that the ones obtained from the study cohort A.

In this chapter, we present the development of two customize scores for the one-year mortality risk of the patients that are admitted in an ICU with a sepsis diagnosis. The objective of this scores is to the allow the stratification of patients into risk groups according to their characteristics of clinical relevance, whereby, the scores would indicate the severity of the condition of the patients.

The predictors that are included in the study cohort B could be which can be roughly grouped in two types, categorical values and continuous numerical values. The categorical values are represented as binary data, and in general they indicate if the patient has a particular characteristic (for instance if the patient is male, or if the patient have hypotension). The continuous numerical variables were divided in groups with different one-year mortality risk for this we found cut-off points (CP) for each of the variables of this type.

Two methodologies were used to find the CP of the continuous variables; the first one finds a single cutoff point that binarized each variables in a group whit high risk and a group with low risk. The second one finds multiple cutoff points for each variable. With each of these methodologies, we developed a score value for each variable in the model, that was calculated as the value of the coefficients in a prediction logistic regression model multiplied by 10 and rounding to the nearest integer. In addition, a constant was added to each integer coefficient to eliminate any negative values. These nonnegative integers are the point values that make up the one-year mortality prediction score for sepsis patients when summed.

Then, the one-year mortality probability was estimated using the score as the only variable in a logistic regression model.

7.2 Methodology

In this section we present the methodology followed to generate the scores for the one-year mortality risk. First we presented a brief description of the study cohort B. Then, we explain in detail the two methodologies used to select the cutoff points of each score predictor, and we present the results of each score.

7.2.1 Dataset

In order to calculate the score we obtained the following predictors from cohort B.

- Routine charted data: The maximum, minimum and mean values during the first 24 hours of the ICU stay of the following vital signs: heart rate, systolic blood pressure, diastolic blood pressure, mean blood pressure, respiratory rate, temperature and peripheral capillary oxygen saturation.
- Laboratory variables: The maximum and minimum values of following laboratory variables obtained from the first 24 hours in the ICU: anion gap, bicarbonate, bilirubin, arterial pH, creatinine, chloride, glucose, hematocrit, platelet count, hemoglobin, lactate, potassium, Partial Thromboplastin Time (PTT), Prothrombin Time (PT), the international normalized ratio (INR), sodium, Blood Urea Nitrogen and White Blood Cell Count (wbc).
- Categorical variables: admission type (elective, urgent, emergency), gender, the receipt of either two treatments (dialysis and mechanical ventilation) and comorbidities according to the Elixhauser Comorbidity groups (30 comorbidities).
- The predictors: admission age, the minimum Glasgow Coma Scale, and the total urinary output over the first 24 hours.

Laboratory measurements and routine charted data were converted into 96 predictors; after that, we used two different methodologies to select the cutoff points of each score predictor. To do that, Study cohort B was randomly divided into two groups: a training subset with 40% of the admissions and a validation subset of 60% of the admissions. Methodologies run over training set. Usually, a training cohort larger than the validation cohort is used, however, after performing multiple tests with different sizes for the training cohort, we found that a sample of 40% is sufficiently representative of our study population, and the performance, as measured by the AUROC, does not improve substantially when larger training subsets (for example 70% or 80%) are used. Figure 7.1, presents the mean AUROC of 100 runs with each training subset percentage and Table 7.1table y presents the improvement obtained from increasing the training subset.

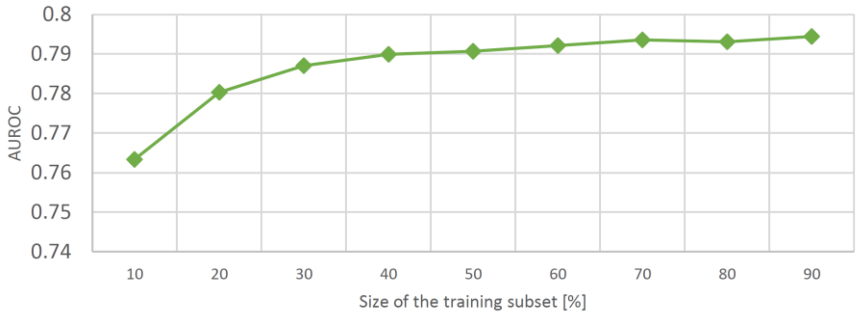


Figure 7.1. AUROC as function of the training subset size.

Table 7.1. Improvement obtained from increasing the training subset

Size of the training subset [%]	AUROC	Improvement
10	0.7633	---
20	0.7803	2.23
30	0.787	0.86
40	0.7899	0.37
50	0.7907	0.1
60	0.7921	0.18
70	0.7935	0.18
80	0.7931	-0.05
90	0.7944	0.16

7.2.2 Binary score

This methodology allow us to obtain a cutoff point (from now CP) for each of the continuous numeric predictor variables; it divide the dataset into two groups based on a set of values for each predictor, those below the CP and those above; afterwards, we calculated the mortality rate of each group and assessed the number of admissions in each group. Thus, the CP was selected taking into account the following criteria:

- Criteria 1: The smallest group contains at least 30% of admissions.
- Criteria 2: The biggest difference between the mortality rates between the groups.

In order to fulfill the criteria 1, we find the 30% and 70% quantiles in the training subset. Then, we generate a regular sequence between those two values with a length of 1.000, each of the elements of the sequence is a candidate cutoff point. We calculate the difference between the mortality rates of the populations over and below the candidate CP, and selected the one that presented a greater difference (criteria 2).

Since the selection of the cutoff points is done over the training subset (6.033 admissions), it varies according to the random split that was made of cohort B; for this reason, we repeat the process 100 times and selected as final CP the mean of the values that presented a bigger difference each time.

To illustrate the process for the cutoff point selection, we present in detail the results for the age and the minimum mean arterial blood pressure; since these variables represent the behavior that variables can suffer when dichotomized.

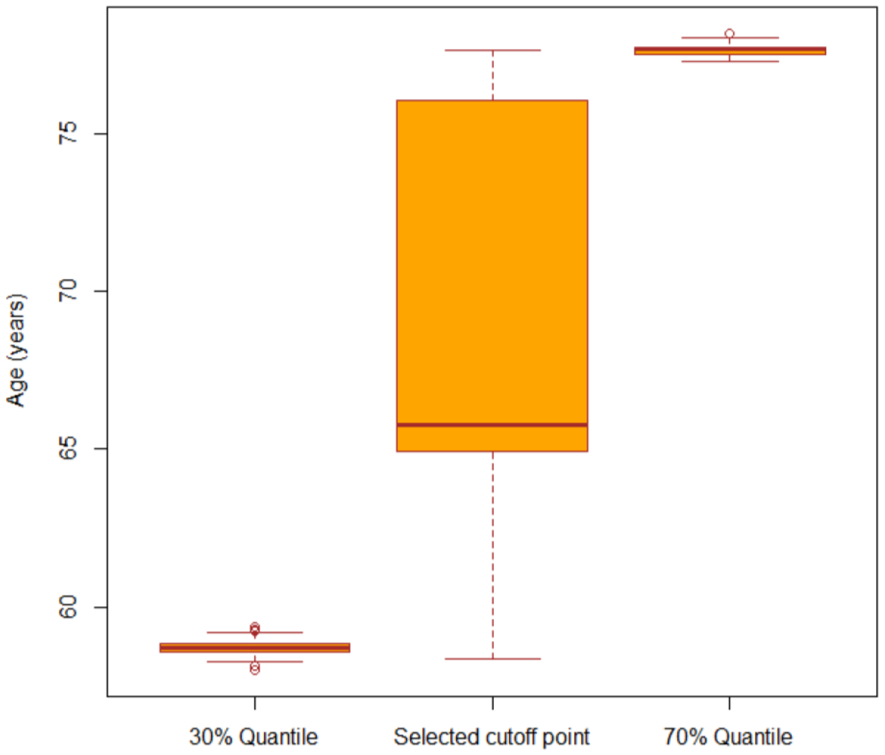


Figure 7.2. Distribution of the 30% quantile, 70% quantile and selected cutoff points for the age over the 100 runs.

Figure 7.2 presents the distribution of the 30% quantiles, 70% quantiles and selected cutoff points for the age in all the 100 runs. It is clear that the selected cutoff point is more sensitive to random divisions than the other two values; however, the average of difference of mortality rates along all the candidate CP does not show much variation since the values change from 18.67391% to 20.66875%, Figure 7.3 presents the average difference for the 100 runs for each candidate CP.

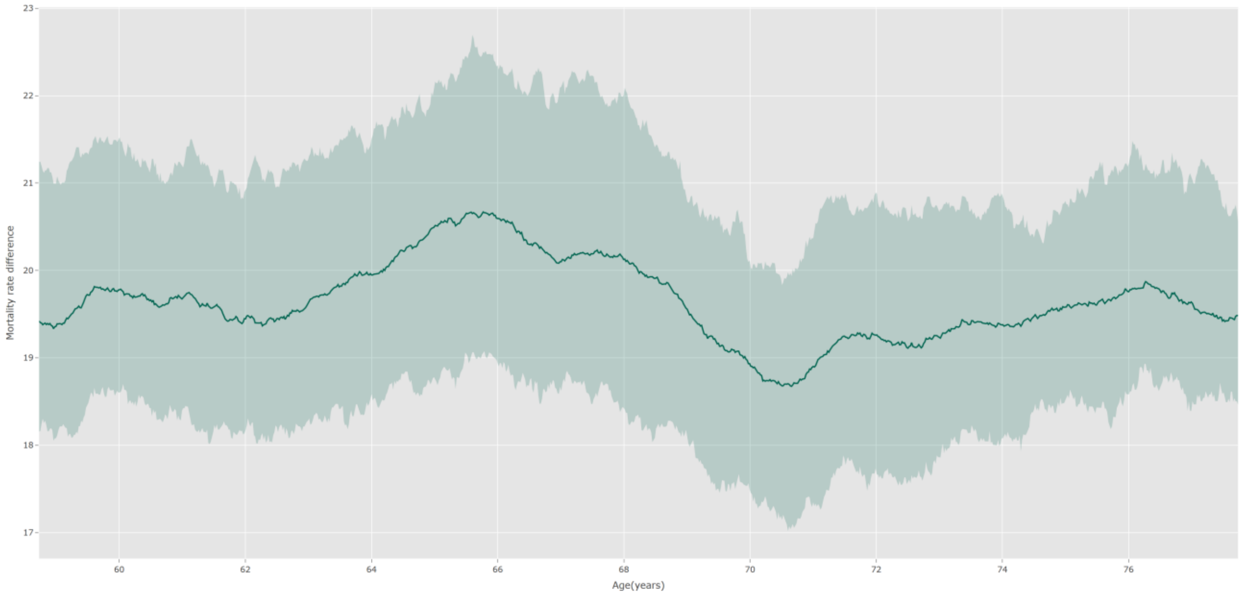


Figure 7.3. Mortality rate difference for the age along the regular sequence between the 30% quantile and 70% quantile values.

Figure 7.4 presents the distribution of the 30% quantiles, 70% quantiles and selected cutoff points for the minimum mean arterial blood pressure in all the 100 runs. In contrast to the age, it can be observed that the selected cutoff point variation over all the runs is small. Figure 7.5 presents the average difference for the mean arterial blood pressure over 100 runs for each candidate CP.

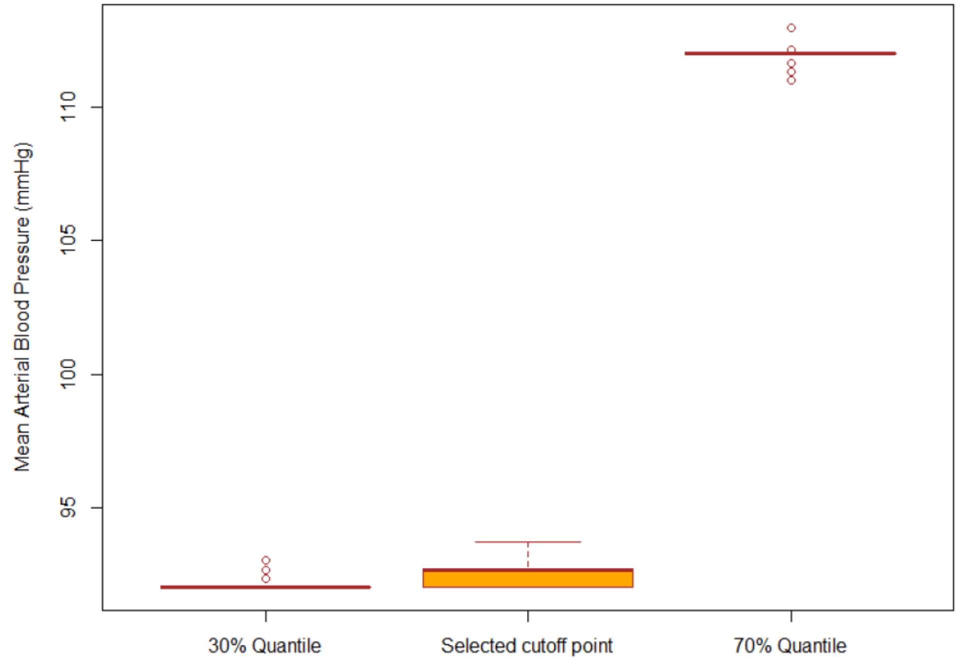


Figure 7.4. Distribution of the 30% quantiles, 70% quantiles and selected cutoff points for the minimum Mean Arterial Blood Pressure over the 100 runs.

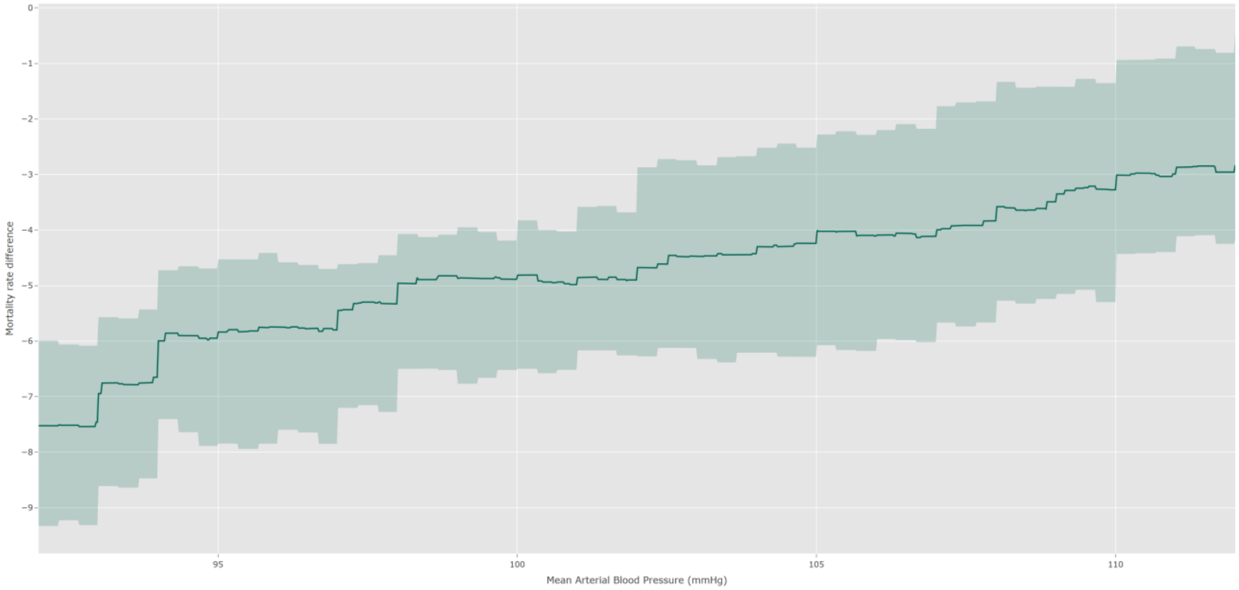


Figure 7.5. Mortality rate difference for the mean arterial blood pressure along the regular sequence between the 30% quantile and 70% quantile values

The mortality rate vs candidate CP (Figure 7.3 and Figure 7.5) show a different behavior for the example variables; for the age are positive which means that the population that are in risk is the one that is over the CP; on the other hand, for the mean arterial blood pressure the differences are negatives which means that the population at risk is the one that is below the CP. According to the CP selection criteria, we look for the biggest difference between the mortality rates of the populations over and below the CP, thus the selected CP is the one with the higher absolute value of the mortality rate difference.

Selected CP allows us binarized all predictors. It means, a one is assigned to each predictor if its value is within the population with a higher mortality rate. After that, we develop a logistic regression (LR) model using this binary data in conjunction with the data taken at the time of ICU admission, the comorbidities and the treatments. The general form of the log-odds, without the intercept coefficient:

$$l = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

In (1) coefficients β_i are the parameters of the model, x_i are predictors and n is the total number of predictors.

In order to obtain point values of the scoring system, the coefficients in (1) were multiplied by 10 and a constant was added to each integer coefficient to eliminate any negative values, then, they were rounded to the nearest integer. After the coefficient transformation:

$$l = (10\beta_1 + \varrho)x_1 + (10\beta_2 + \varrho)x_2 + \cdots + (10\beta_n + \varrho)x_n \quad (2)$$

In (2) ϱ is a constant added to coefficient to eliminate the negative values $and\ is$ equivalent to the smallest of the coefficients multiplied by ten. Rearranging the term, we got that:

$$l = (10\beta_1 x_1 + 10\beta_2 x_2 + \cdots + 10\beta_n x_n) + \varrho(x_1 + x_2 + \cdots + x_n) \quad (3)$$

The transformation process change the log-odds in two ways, of which, the multiplication of 10 with each of the coefficients β_i do not affect the performance of the model, however the term $\varrho(x_1 + x_2 + \cdots + x_3)$ could add important variations to the log odds, and it effect increases with the number of predictors; for this reason the score is defined as follows:

$$Score = (\beta_1^T x_1 + \beta_2^T x_2 + \cdots + \beta_n^T x_n) - \varrho(x_1 + x_2 + \cdots + x_n) \quad (4)$$

Where the coefficients $\beta_i^T$are the β_i coefficients multiplied by ten and rounded, the rounding of this coefficients is done in favor of clinical operationalization and score interpretability and it does not significantly affect the final performance.

We executed multiple runs for different training subsets (in each run we randomly selected 40% of the study cohort B as training subset), obtaining different discrimination and calibration performance; and since the score is based on logistic regression, the coefficients are dependent to the selected training subset used to construct it; according to this, we present the score parameters obtained with a particular training subset that reported a discrimination performance close to the mean of the AUROC obtained with the all the runs and presented an adequate calibration.

The score development methodology presented above, brings back a set of cutoff point that divides the population into two groups, the difference in the one-year mortality rate between the two groups indicates how good such a predictor is. For this reason, different thresholds for said mortality rate difference were evaluated; with an increase in the one-year mortality rate difference threshold, the number of variables and the performance of the model decrease, making the score simpler but less discriminative. Figure 7.6 presents the behavior of the performance of the model according to the one-year mortality rate difference threshold.

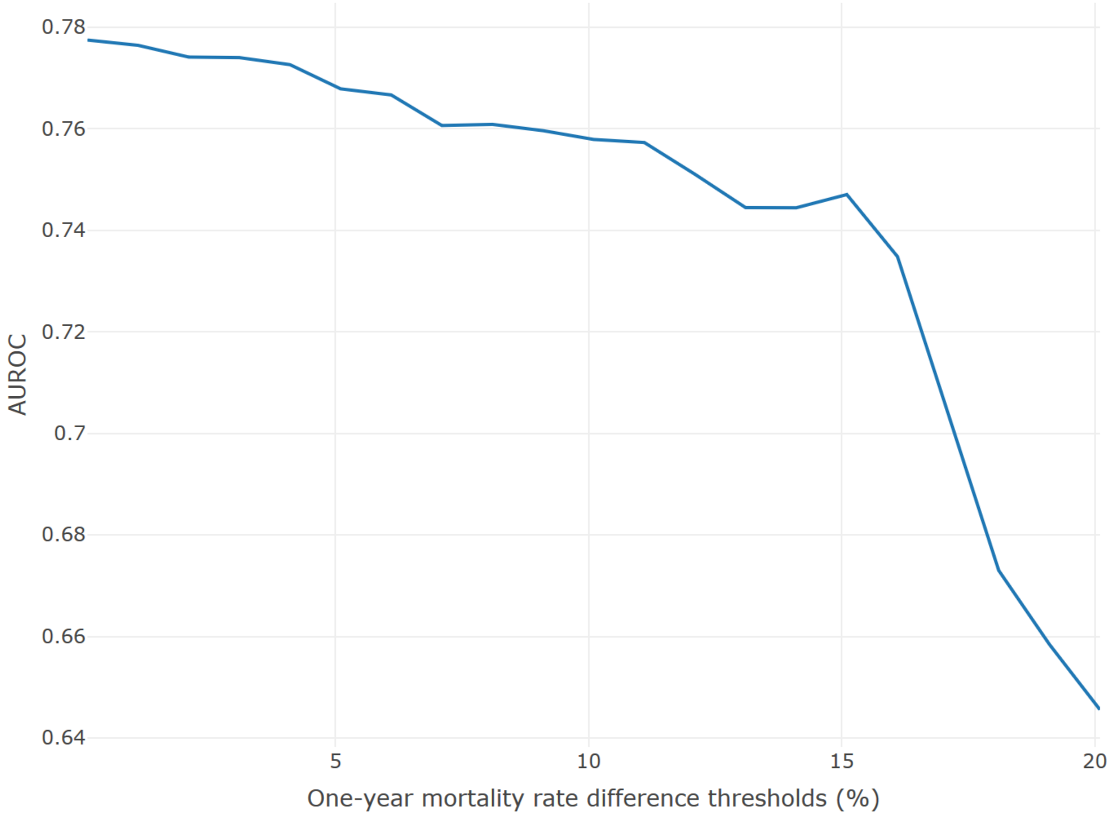


Figure 7.6. Different AUROC for One-year mortality rate difference thresholds.

The coefficients for the one-year mortality prediction score for ICU sepsis patients, the selected cutoff points (CP), the mortality rates for the population below and above the CP and the point value for each of the predictor variables that present a mortality rate difference greater than 5 are shown in Table 7.2.

The coefficients and mortality rates for the comorbidities, treatment and demographic variables are presented in Table 7.3. With the coefficients presented in Table 7.2 and Table 7.3 the score for the one-year mortality prediction of sepsis patients is calculated as follows:

$$Score = (\beta_1^T x_1 + \beta_2^T x_2 + \cdots + \beta_n^T x_n) - 138(x_1 + x_2 + \cdots + x_n)$$

Where β_i^T are the coefficients presented in Table 7.2 and Table 7.3, and x_i are one if the patient is in the group with higher mortality rate for each variable and zero otherwise. For instance, if the patient admission age is greater than 73.9 it associated x will be 1, and if the patient minimum diastolic blood pressure over the first 24 hours of ICU admission is greater than 41 mmHg it associated x will be 0.

Table 7.2. Scoring System for the One-Year Mortality Prediction of Sepsis Patients in Intensive Care Units.

Predictor	Mortality rate below CP	Mortality rate above CP	Cutoff point (CP)	Coefficient
Admission age	35	55	73.91037	145
Anion gap min	39	52	14	138
Anion gap max	40	51	18.8	139
creatinine min	38	53	1.431111	138
creatinine max	36	50	1.39	136
hemoglobin max	51	37	10.8702	141
lactate min	40	50	1.622424	142
Partial thromboplastin time min	38	53	32.9498	141
Partial thromboplastin time max	38	50	36.71332	140
International Normalized Ratio min	38	57	1.4	138
International Normalized Ratio max	38	54	1.632727	139
Prothrombin Time min	37	55	14.95947	141
Prothrombin Time max	38	53	16.70015	135
Blood Urea Nitrogen min	35	56	29.76667	142
Blood Urea Nitrogen max	32	53	28.28081	142
Urine Output	52	36	1392.887	141
Diastolic blood pressure min	49	36	41.00909	140
Mean blood pressure min	50	37	53.7633	139
Mean blood pressure avg	47	36	78.75689	137
Temperature min	48	35	36.30914	140

The final score possible maximum number is 85 (however no patient gets the maximum score in the dataset). The probability of one-year mortality on the validation subset was estimated using the final score as the sole variable in a logistic regression model, model discrimination was examined and the obtained AUROC of 0.768 (95% Confidence Interval: 0.761 - 0.778). To access the calibration of the score, Hosmer–Lemeshow test was used and a value of 0.9 indicating that there is no evidence of poor fit.

The admissions of the validation subset were divided into ten equal size groups according to the increasing estimated probabilities of one-year mortality given by the model, so that in the first group are those admissions that have the lowest probabilities of dying, and in the last group are those admissions with the highest probabilities of dying. For each group the observed and the estimated number of deaths were calculated and compared graphically and presented in Figure 7.7.

Table 7.3. Comorbidities treatments and demographics variables for the Scoring System for the One-Year Mortality Prediction of Sepsis Patients in Intensive Care Units

Predictor	Mortality rate without presence	Mortality rate with presence	Coefficient
Admission type: EMERGENCY	27	44	143
Gender: male	42	44	140
Congestive heart failure	38	50	141
Cardiac arrhythmias	38	53	140
Valvular disease	42	46	137
Pulmonary circulation	43	46	139
Peripheral vascular	42	47	141
Hypertension	43	43	134
Paralysis	43	40	138
Other neurological	43	41	141
Chronic pulmonary	42	45	141
Diabetes uncomplicated	43	43	137
Diabetes complicated	43	43	136
Hypothyroidism	43	46	139
Renal failure	40	51	140
Liver disease	42	50	144
AIDS	43	45	139
Lymphoma	42	69	145
Metastatic cancer	40	82	160
Solid tumor	42	59	146
Rheumatoid arthritis	43	42	136
Coagulopathy	40	52	138
Obesity	44	32	135
Weight loss	43	48	139
Fluid electrolyte	40	46	140
Blood loss anemia	43	46	139
Deficiency anemias	43	43	136
Alcohol abuse	44	31	136
Drug abuse	44	21	134
Psychoses	44	30	134
Depression	44	35	136
Renal replacement therapy (RRT)	42	56	143

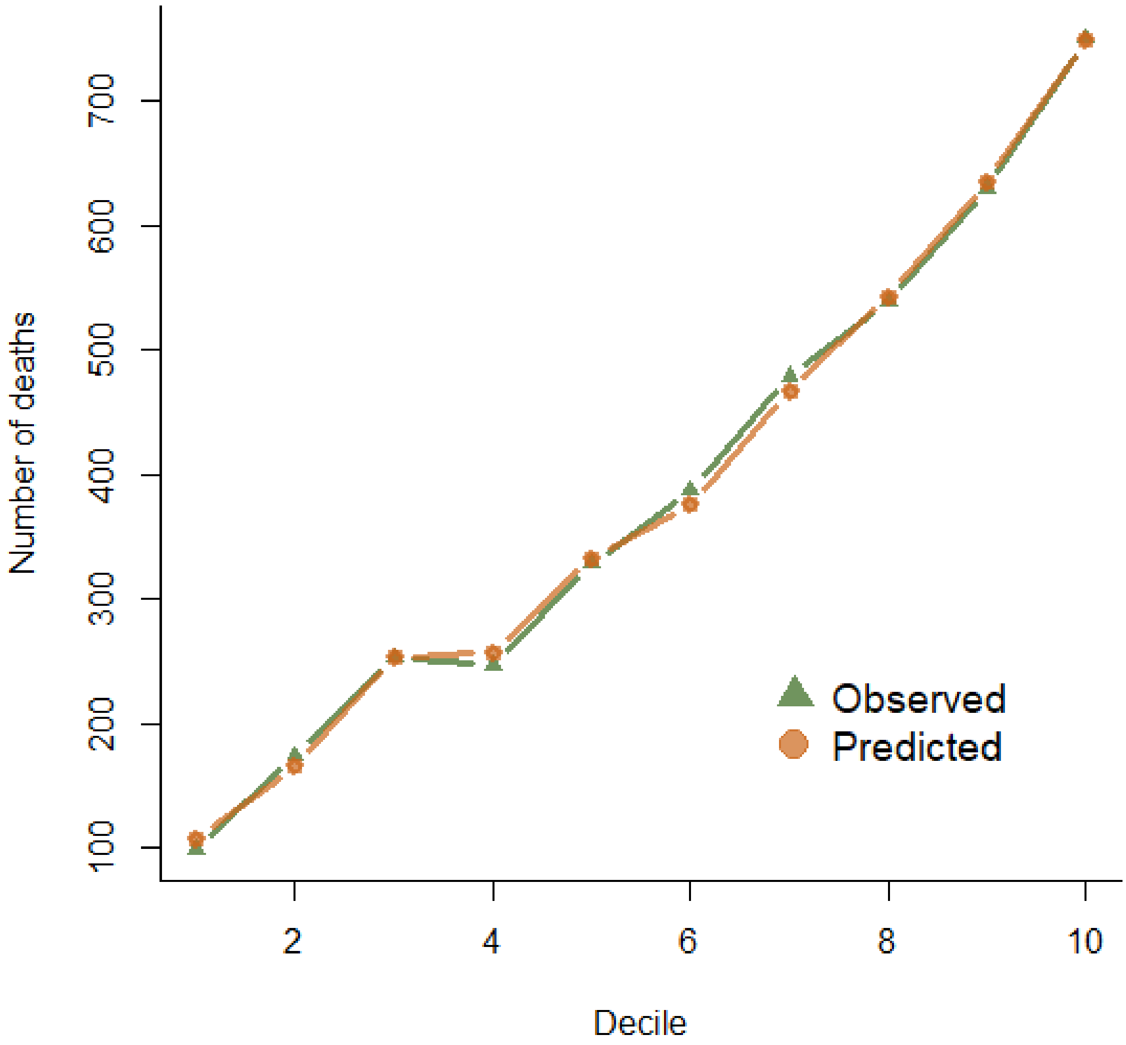


Figure 7.7. Comparison of observed versus predicted number of deaths by groups of increasing probability of one-year mortality.

To benchmark the scoring system, the AUROC of 10 adjusted Severity of Illness Scores on the validation subset were calculated over 100 runs with different population partitions, the results are presented in Figure 7.8. In order to evaluate the calibration of the score, we check how the Hosmer-Lemeshow test performs in the said 100 repeated samples, the we calculated the proportion of p-values which are less than 0.05. From 100 runs, the Hosmer-Lemeshow test gave a significant p-value, indicating poor fit, on only 2% of occasions.

The AUROC analysis presented in Figure 7.8, ratifies that a the development of entirely new models, that incorporate additional variables that enhance discrimination performance. However when analyzing in detail the information in Table 7.2 it is clear that variables that present a high different between the one-year mortality rates of the populations below and above the CP do no always report a higher score point; for instance the maximum Prothrombin Time reports a difference between mortality rates of 15%, but the score point assigned of 135 is among the smallest; on the other hand, the minimum lactate has a lower difference between mortality rates but it associated score point is higher (142). The mentioned above indicates that for some variables there is a more complex relation between the cutoff points and the mortality rates, as example, one could think in the temperature, a physiological parameter that is pathologic both above (hyperthermia) and below (hypothermia) a CP which suggest that for parameter associate with the temperature it could be better to have more than 2 groups. In the next section we present the multiple cutoff points score, an approach that seeks to implement the aforementioned.

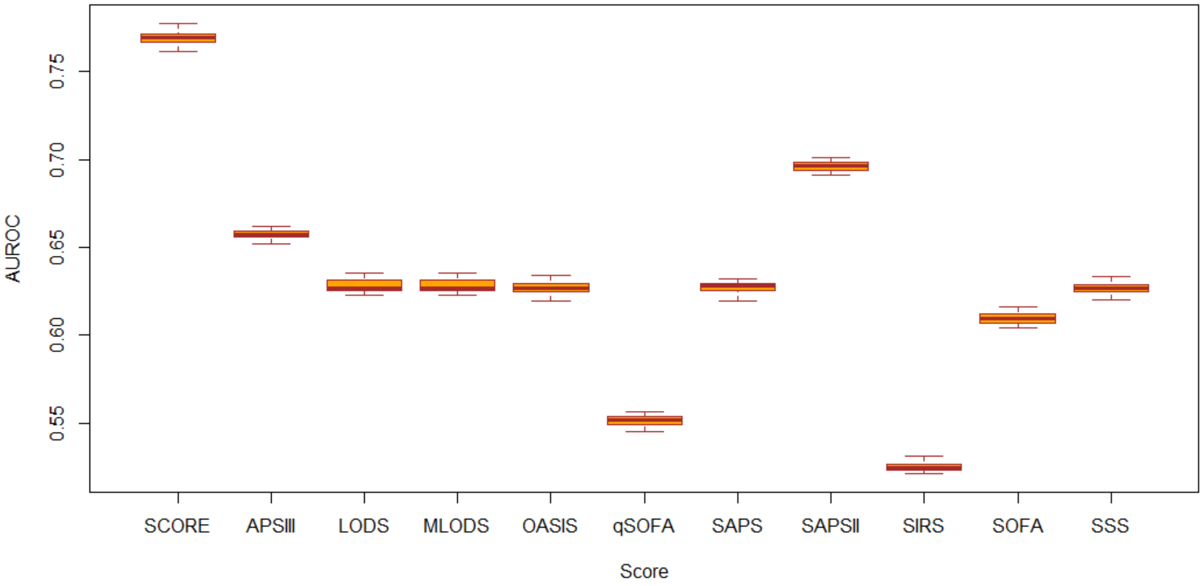


Figure 7.8. AUROC comparison for the developed score and ten scoring systems.

7.2.3 Multiple cutoff points score

The score presented in previous section dichotomizes each continues predictor, and a subset of the population with higher risk can be found; however, it is possible that within the same variable there are several groups with different mortalities, this approach seeks to find the cutoff points that allow the identification of said groups and shows the improvement with respect to the binary score.

In this methodology, a multiple cutoff points were obtained for each of the continuous numeric predictor variables. For this, we selected the unique cutoff points from the minimum and maximum values along with the deciles that divide the range of each variable into continuous intervals with equal probabilities.

Then we computed the one-year mortality rate of each of the groups formed by the unique cutoff points and use agglomerative hierarchical cluster analysis (HCA) with a Euclidian measure of dissimilarity between each one-year mortality group rate.

After that, we divide each of the continuous numeric predictors in the training subset into discrete values according to the groups obtained with the HCA, and calculated the one-year mortality rate for the new groups and for each variables we found the group with the lowest mortality rate and set it as references. The remaining groups were used along with conjunction with the data taken at the time of ICU admission, the comorbidities and the treatment information to generate a score with the same methodology presented in the previous section.

To illustrate the process for the cutoff point finding, we present in detail the results for the age, the minimum mean arterial blood pressure and the glucose minimum.

The minimum, maximum and the deciles that divide the range of the admission age are presented in Table 7.4. These values divides the population according to the age in 10 different groups, with different mortality rates, which are shown in Figure 7.9 It can be observed that as expected, in general, the one-year mortality rate increases with the admission age of the patients; however, there are two details that indicate that there is no need for all these cutoff points: first the last group (patients between 86.6 years old and 90 years old) shows a lower mortality rate than the ninth group (patients from 81.8 years old to 86.6 years old); and second, the third and fourth groups present a similar one-year mortality rate, indicating that patients older than 52.3 years old and younger than 63.3 years old have the same risk.

Table 7.4. Candidate cutoff points and reference values for the admission age.

	Candidate Cutoff points									
Minimum	10%	20%	30%	40%	50%	60%	70%	80%	90%	Maximum
17.0187	44.20862	52.34046	58.24704	63.33326	68.1763	73.28082	77.76584	81.87388	86.61214	90

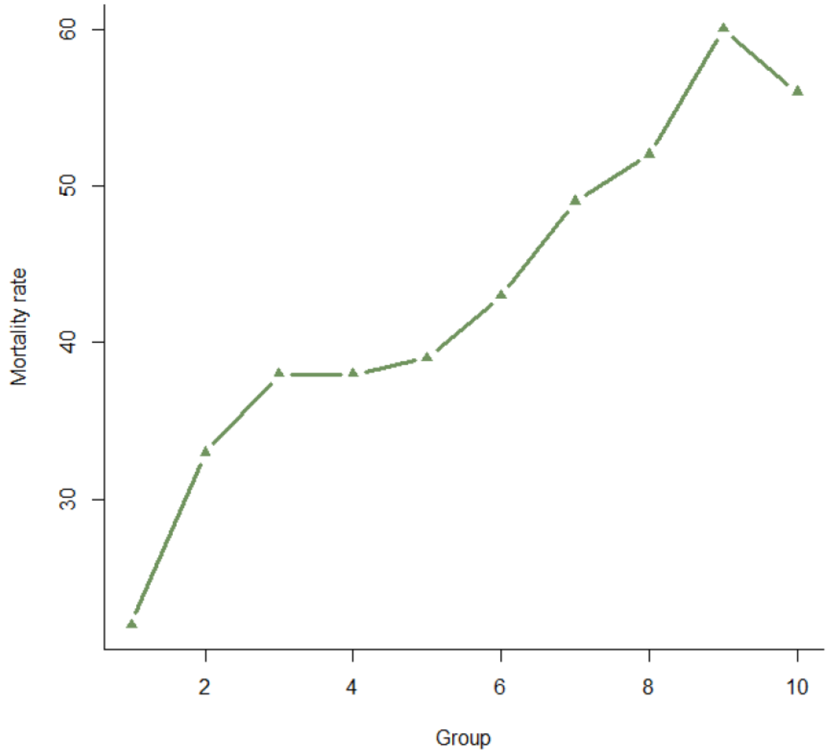


Figure 7.9. One-year mortality rate by groups according to the admission age.

According to the above, an agglomerative hierarchical cluster analysis with a Euclidian measure of dissimilarity between the groups was used to select the final cutoff points for the admission age. Figure 7.10 presents the dendogram of the process. It can be observed that cutting the dendogram at the height of two will result in five groups (presented in Figure 7.11) from which, the first group will be considered the reference, meaning that they patients in this group are considered to have the smallest risk.

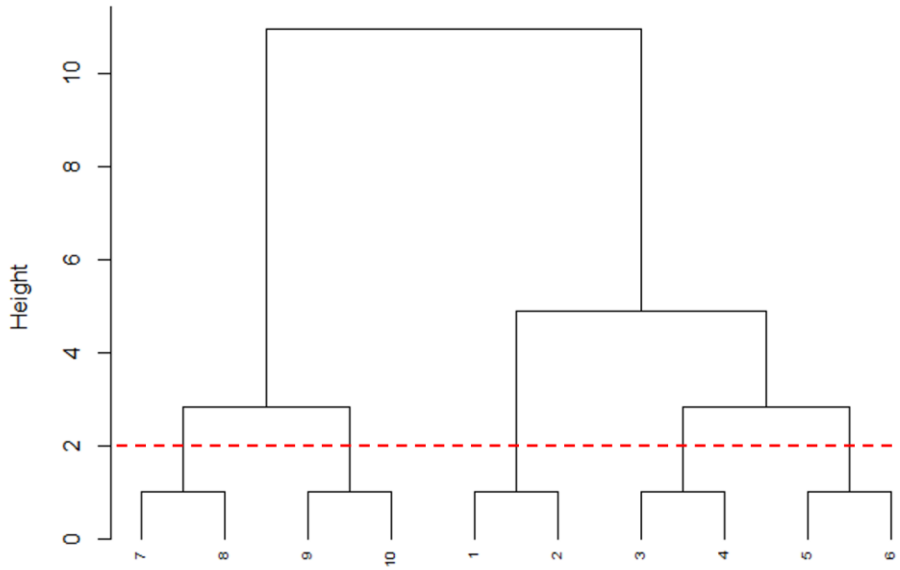


Figure 7.10. Agglomerative hierarchical cluster analysis for the admission age groups.

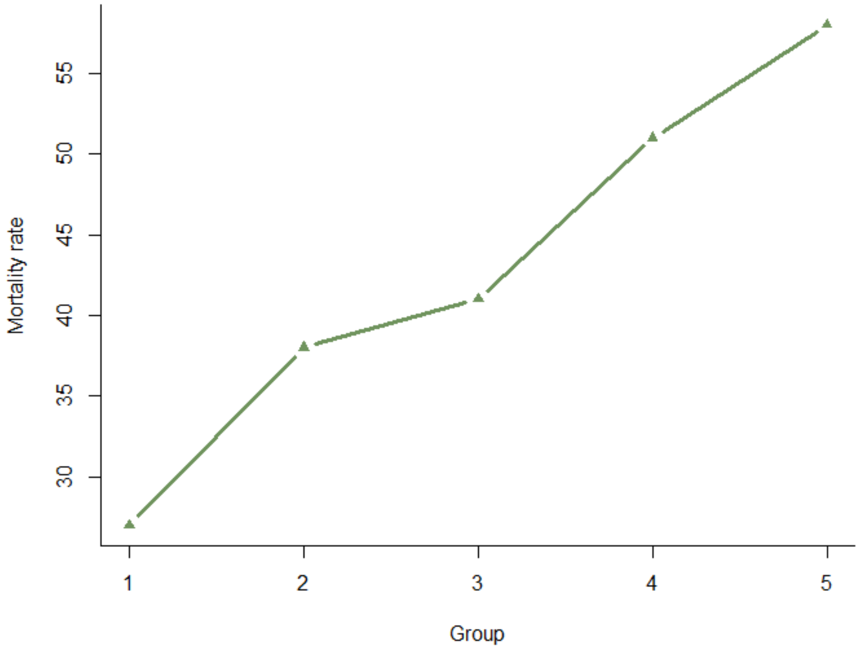


Figure 7.11. One-year mortality rate for the final groups of the admission age.

The minimum, maximum and the deciles that divide the range of the minimum mean blood pressure are presented in Table 7.5. This values divides the population according to the minimum mean blood pressure in 10 different groups, with different mortality rates that decrease when the values of minimum mean blood pressure increase (Figure 7.12 A). After the agglomerative hierarchical cluster analysis (Figure 7.12

C), we found out four groups with the one-year mortality rates presented in Figure 7.12 B. It can be observed that for the minimum mean blood pressure the reference group is the last one.

Table 7.5. Candidate cutoff points and reference values for the minimum mean blood pressure.

	Candidate Cutoff points									
Minimum	10%	20%	30%	40%	50%	60%	70%	80%	90%	Maximum
0.79	38.6667	45.333	49	52	55	57	60	64	70	119

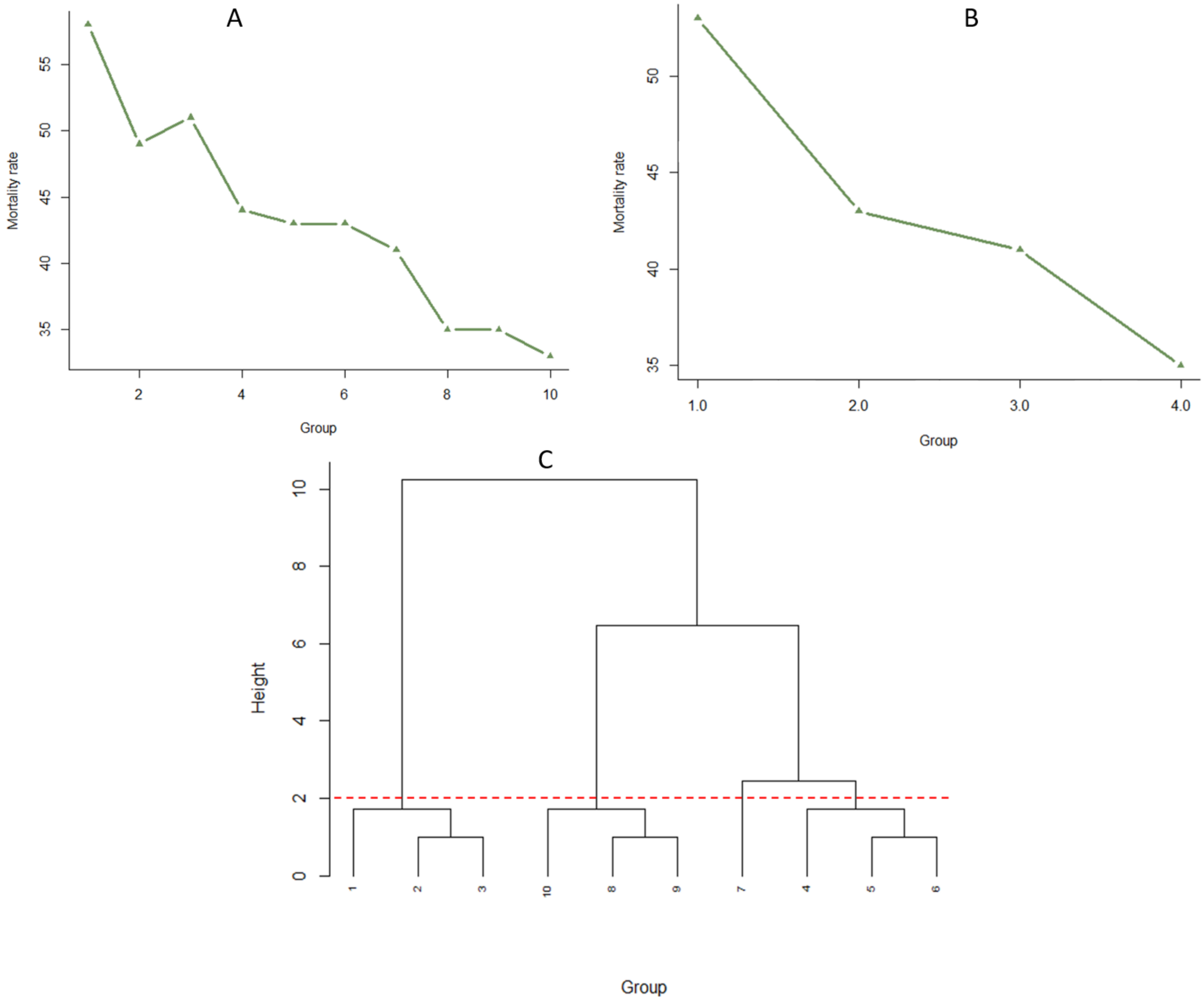


Figure 7.12. Cutoff selection process for the minimum mean arterial blood pressure.

The minimum, maximum and the deciles that divide the range of the minimum glucosee are presented in Table 7.6. These values divides the population according to the minimum mean blood pressure in 10 different groups (Figure 7.13 A). After the agglomerative hierarchical cluster analysis (Figure 7.13 C), we found out four groups with the one-year mortality rates presented in Figure 7.13 B. It can be observed that for the minimum glucose the reference group is the third one.

Table 7.6. Candidate cutoff points and reference values for the minimum glucose.

	Candidate Cutoff points									
Minimum	10%	20%	30%	40%	50%	60%	70%	80%	90%	Maximum
0.85	67	80	88	96	102	109	118	129	147	410

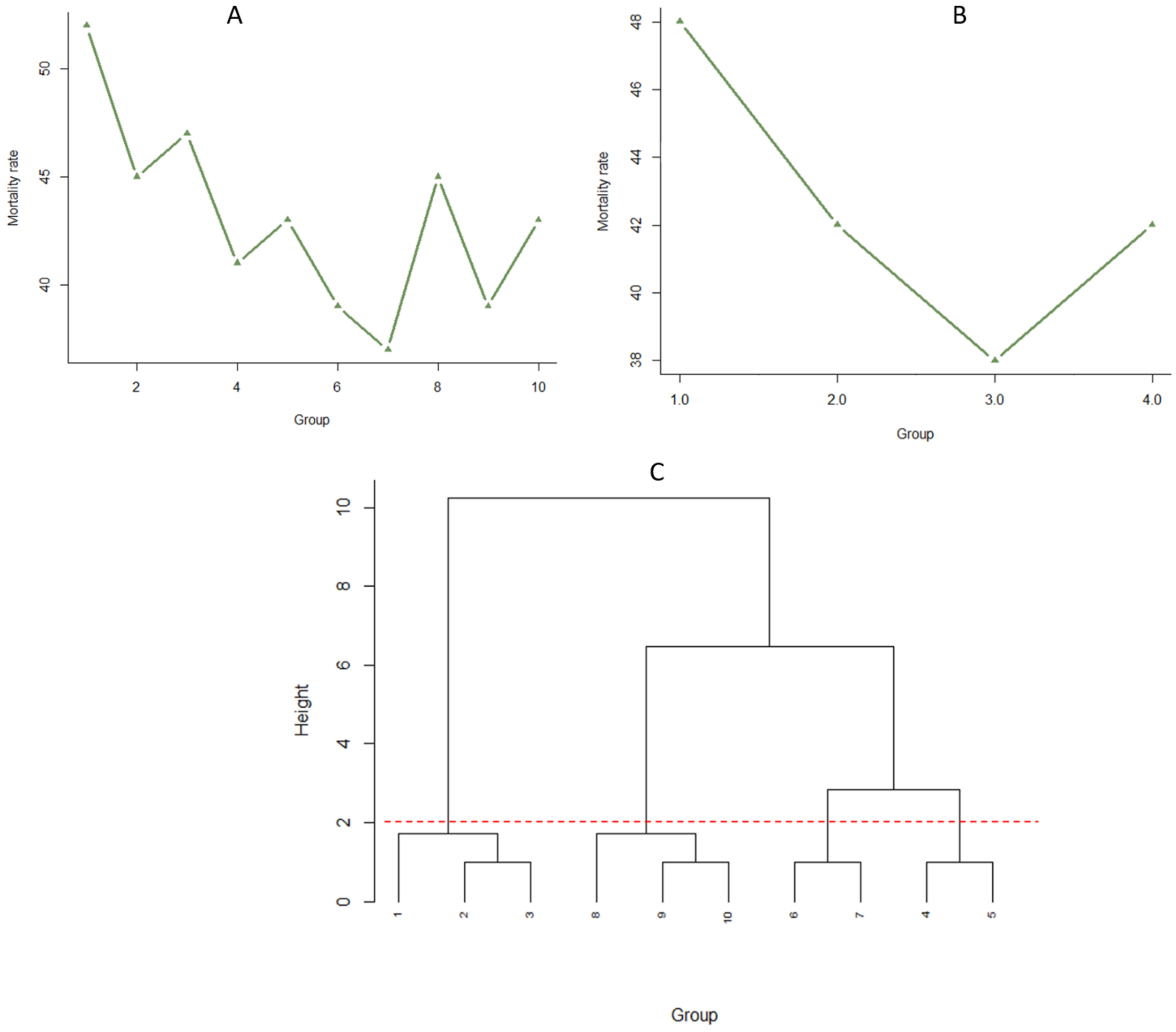
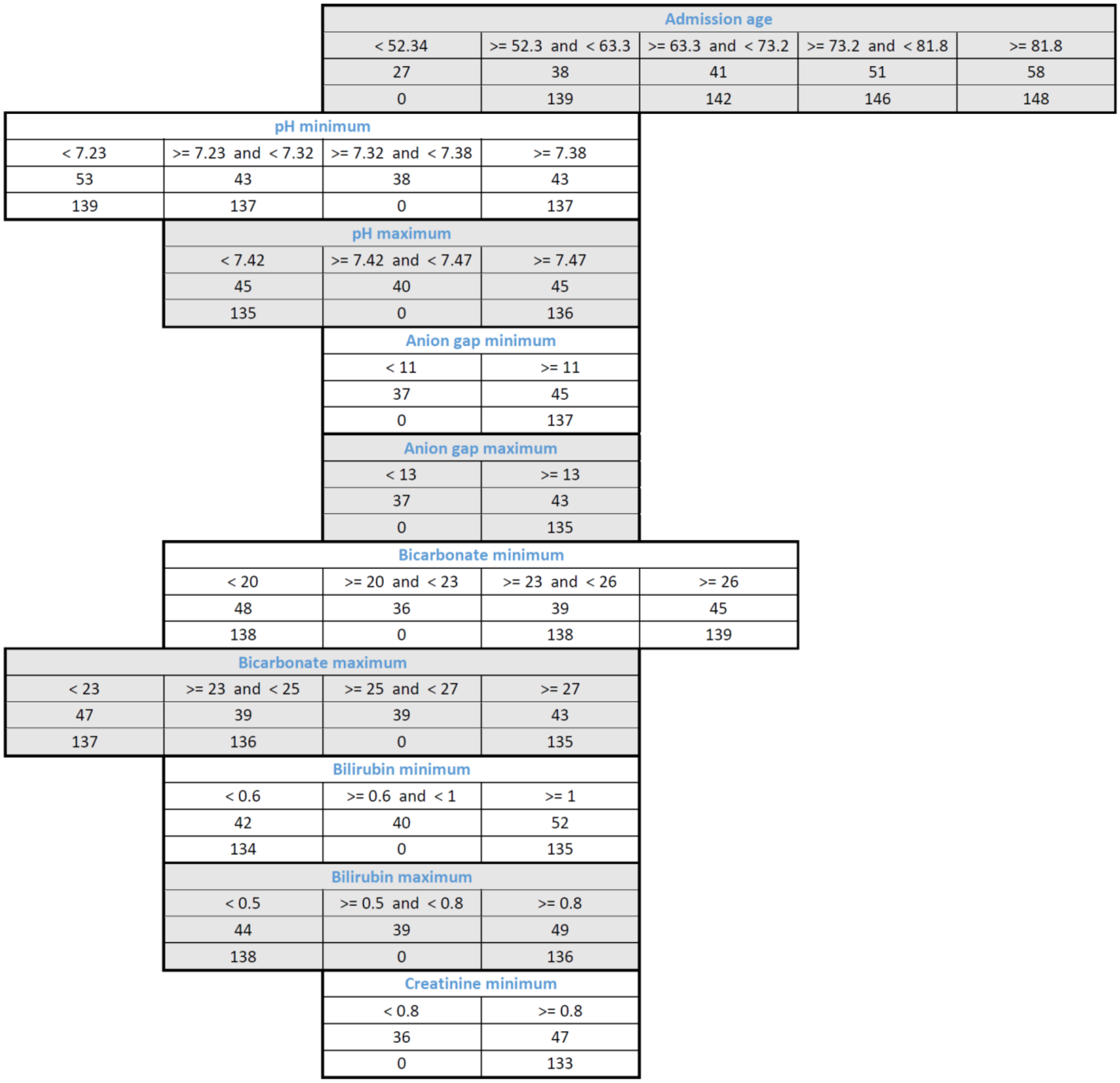


Figure 7.13. Cutoff selection process for the minimum glucose.

The coefficients for the one-year mortality prediction score for ICU sepsis patients, the selected cutoff points (CP) for each predictor, the mortality rates among the groups and the point value for each of the predictor variables are shown in Table 7.7. The coefficients and mortality rates for the comorbidities, treatment and demographic variables are presented in Table 7.8**.** The score for the one-year mortality prediction of sepsis patients is calculated as follows:

$$Score = (\beta_1^T x_1 + \beta_2^T x_2 + \cdots + \beta_n^T x_n) - 135(x_1 + x_2 + \cdots + x_n)$$

Table 7.7. One-year mortality prediction score for sepsis patients within the ICU. Each block represents one of the predictors, the columns of such block represent the number of groups for a particular predictor; the first row of each block presents the name of the predictor, the second row indicates the ranges that defines each group, the third row shows the mortality rate of each group, and the fourth row present the score point values β^T*. Each of the block have a column with a score point value of 0, which is the reference group for the particular predictor.*



Admission age				
< 52.34	>= 52.3 and < 63.3	>= 63.3 and < 73.2	>= 73.2 and < 81.8	>= 81.8
27	38	41	51	58
0	139	142	146	148

pH minimum			
< 7.23	>= 7.23 and < 7.32	>= 7.32 and < 7.38	>= 7.38
53	43	38	43
139	137	0	137

pH maximum		
< 7.42	>= 7.42 and < 7.47	>= 7.47
45	40	45
135	0	136

Anion gap minimum	
< 11	>= 11
37	45
0	137

Anion gap maximum	
< 13	>= 13
37	43
0	135

Bicarbonate minimum			
< 20	>= 20 and < 23	>= 23 and < 26	>= 26
48	36	39	45
138	0	138	139

Bicarbonate maximum			
< 23	>= 23 and < 25	>= 25 and < 27	>= 27
47	39	39	43
137	136	0	135

Bilirubin minimum		
< 0.6	>= 0.6 and < 1	>= 1
42	40	52
134	0	135

Bilirubin maximum		
< 0.5	>= 0.5 and < 0.8	>= 0.8
44	39	49
138	0	136

Creatinine minimum	
< 0.8	>= 0.8
36	47
0	133

Table 7.7 (Continued). One-year mortality prediction score for sepsis patients within the ICU.

Creatinine maximum		
< 0.8	>= 0.8 and < 1.2	>= 1.2
36	35	47
136	0	135

Chloride minimum			
< 98	>= 98 and < 102	>= 102 and < 105	>= 105
48	42	39	42
138	136	0	136

Chloride maximum		
< 109	>= 109 and < 113	>= 113
45	37	43
140	0	136

Hematocrit minimum	
< 33.7	>= 33.7
44	37
134	0

Hematocrit maximum	
< 39.3	>= 39.3
44	37
134	0

Hemoglobin minimum			
< 8.5	>= 8.5 and < 10	>= 10 and < 11.3	>= 11.3
50	45	39	34
141	138	136	0

Hemoglobin maximum			
< 10.8	>= 10.8 and < 11.8	>= 11.8 and < 13.1	>= 13.1
49	43	39	34
142	140	137	0

Lactate minimum		
< 1.2	>= 1.2 and < 1.7	>= 1.7
36	40	52
0	138	140

Table 7.7 (Continued). One-year mortality prediction score for sepsis patients within the ICU.

Lactate maximum	
< 1.2	>= 1.2
35	45
0	135

Platelet minimum			
< 104	>= 104 and < 183	>= 183 and < 281	>= 281
53	40	40	42
138	134	0	135

Platelet maximum		
< 235	>= 235 and < 302	>= 302
44	39	42
136	0	135

Potassium minimum		
< 3.6	>= 3.6 and < 4	>= 4
41	38	49
137	0	136

Potassium maximum			
< 4	>= 4 and < 4.4	>= 4.4 and < 5.4	>= 5.4
39	38	44	50
137	0	137	138

Partial Thromboplastin Time (PTT) minimum		
< 27	>= 27 and < 33.1	>= 33.1
35	40	55
0	135	139

Partial Thromboplastin Time (PTT) maximum		
< 27.9	>= 27.9 and < 37.3	>= 37.3
34	39	51
0	135	136

International Normalized Ratio (INR) minimum			
< 1	>= 1 and < 1.3	>= 1.3 and < 1.6	>= 1.6
39	36	45	57
136	0	135	137

International Normalized Ratio (INR) maximum		
< 1.2	>= 1.2 and < 1.5	>= 1.5
34	39	50
0	137	138

Prothrombin Time (PT) minimum			
< 12.8	>= 12.8 and < 13.7	>= 13.7 and < 15.2	>= 15.2
36	36	41	54
0	134	136	137

Prothrombin Time (PT) maximum	
< 13.4	>= 13.4
34	45
0	134

Sodium minimum		
< 136	>= 136 and < 139	>= 139
43	38	45
135	0	138

Sodium maximum		
< 140	>= 140 and < 144	>= 144
44	39	47
136	0	137

Table 7.7 (Continued). One-year mortality prediction score for sepsis patients within the ICU.

Blood Urea Nitrogen minimum			
< 16	>= 16 and <28	>= 28 and <44	>= 44
29	39	53	59
0	137	140	142

Blood Urea Nitrogen maximum			
< 16	>= 16 and <25	>= 25 and <36	>= 36
28	32	43	55
0	135	136	137

White blood cell count minimum		
< 8.4	>= 8.4 and < 12.5	>= 12.5
44	40	45
136	0	136

White blood cell count maximum		
< 10.4	>= 10.4 and < 13.5	>= 13.5
44	38	43
137	0	136

Urine output			
< 1035	>= 1035 and < 1805.2	>= 1805.2 and < 2775	>= 2775
57	44	34	29
142	139	137	0

Heart rate minimum		
< 72	>= 72 and < 86	>= 86
43	41	44
135	0	135

Heart rate maximum			
< 96	>= 96 and < 108	>= 108 and < 128	>= 128
42	38	43	49
134	0	137	138

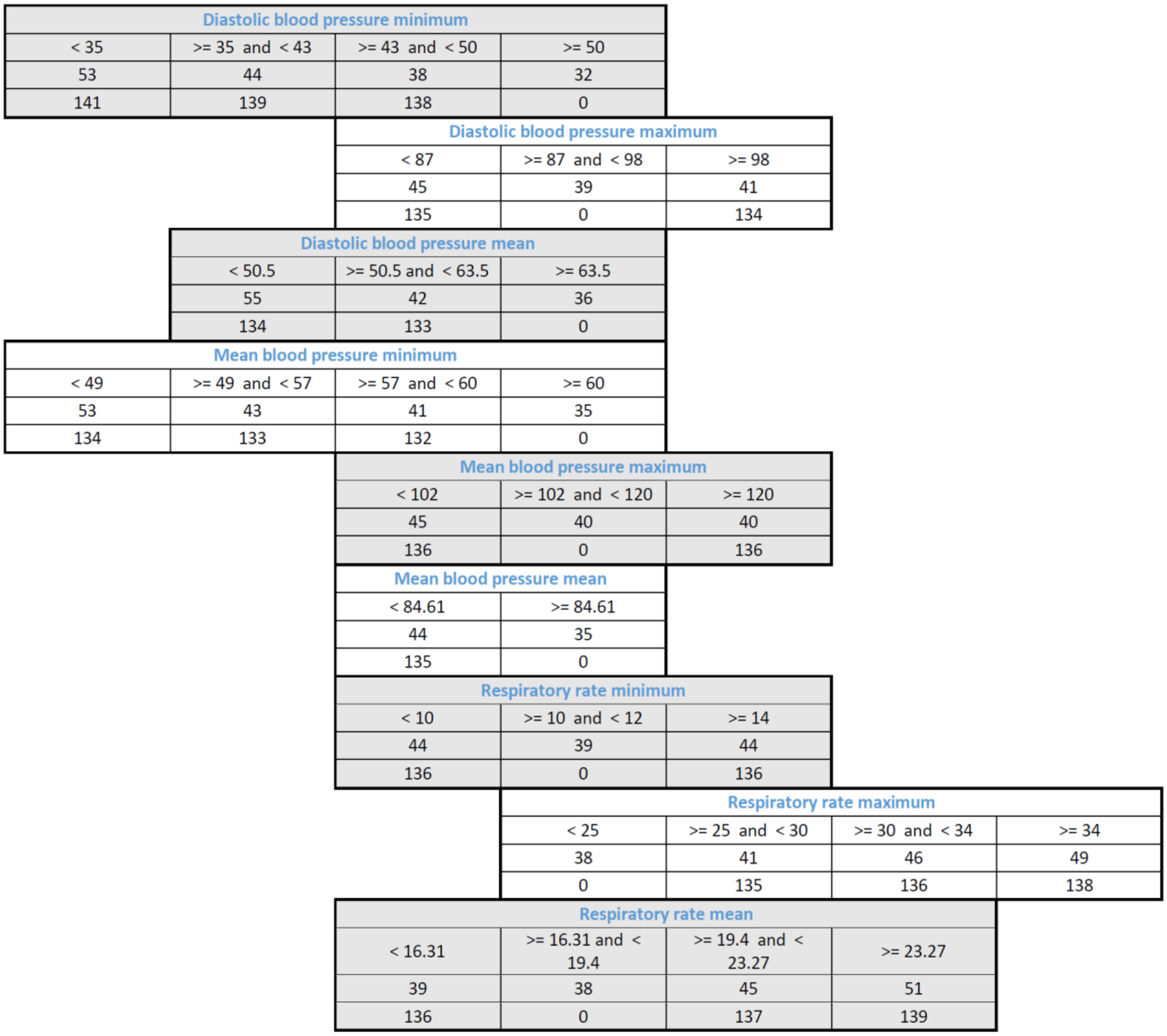
Heart rate mean			
< 83.52	>= 83.52 and < 91.88	>= 91.88 and < 102.45	>= 102.45
42	41	43	46
137	0	135	136

Systolic blood pressure minimum		
< 78.6	>= 78.6 and < 100	>= 100
53	41	35
138	137	0

Systolic blood pressure maximum	
< 161	>= 161
44	39
136	0

Systolic blood pressure mean		
< 109.7	>= 109.7 and < 130.3	>= 130.3
50	39	37
134	133	0

Table 7.7 (Continued). One-year mortality prediction score for sepsis patients within the ICU.



Diastolic blood pressure minimum			
< 35	>= 35 and < 43	>= 43 and < 50	>= 50
53	44	38	32
141	139	138	0

Diastolic blood pressure maximum		
< 87	>= 87 and < 98	>= 98
45	39	41
135	0	134

Diastolic blood pressure mean		
< 50.5	>= 50.5 and < 63.5	>= 63.5
55	42	36
134	133	0

Mean blood pressure minimum			
< 49	>= 49 and < 57	>= 57 and < 60	>= 60
53	43	41	35
134	133	132	0

Mean blood pressure maximum		
< 102	>= 102 and < 120	>= 120
45	40	40
136	0	136

Mean blood pressure mean	
< 84.61	>= 84.61
44	35
135	0

Respiratory rate minimum		
< 10	>= 10 and < 12	>= 14
44	39	44
136	0	136

Respiratory rate maximum			
< 25	>= 25 and < 30	>= 30 and < 34	>= 34
38	41	46	49
0	135	136	138

Respiratory rate mean			
< 16.31	>= 16.31 and < 19.4	>= 19.4 and < 23.27	>= 23.27
39	38	45	51
136	0	137	139

Table 7.7 (Continued). One-year mortality prediction score for sepsis patients within the ICU.

Temperature minimum	
< 36.67	>= 36.67
46	32
137	0

Temperature maximum	
< 38.39	>= 38.39
45	34
136	0

Temperature mean			
< 36.31	>= 36.31and < 36.67	>= 36.67 and < 37.18	>= 37.18
58	47	40	33
143	140	138	0

Peripheral capillary oxygen saturation minimum				
< 88	>= 88 and < 91	>= 91 and < 93	>= 93 and < 95	>= 95
52	46	40	38	40
139	138	137	0	137

Peripheral capillary oxygen saturation mean	
< 98.93	>= 98.93
44	42
135	0

Glucose minimum			
< 88	>= 88 and < 102	>= 102 and < 118	>= 118
48	42	38	42
137	136	0	135

Glucose maximum	
< 125	>= 125
41	44
0	138

Glucose mean			
< 117.26	>= 117.26 and < 133.8	>= 133.8 and < 155.5	>= 155.5
45	40	42	44
135	0	136	137

Table 7.8. Comorbidities treatments and demographics variables for the Scoring System for the One-Year Mortality Prediction of Sepsis Patients in ICU.

Predictor	Mortality rate without presence	Mortality rate with presence	Coefficient
Admission type: EMERGENCY	32	43	141
Gender: Male	42	44	138
Congestive heart failure	38	51	138
Cardiac arrhythmias	39	51	137
Valvular disease	42	45	134
Pulmonary circulation	43	46	136
Peripheral vascular	42	47	138
Hypertension	45	41	132
Paralysis	43	35	135
Other neurological	43	42	139
Chronic pulmonary	42	46	136
Diabetes uncomplicated	43	43	135
Diabetes complicated	43	41	134
Hypothyroidism	42	47	136
Renal failure	40	51	137
Liver disease	41	54	142
AIDS	43	39	137
Lymphoma	43	56	141
Metastatic cancer	40	80	159
Solid tumor	42	58	144
Rheumatoid arthritis	43	39	135
Coagulopathy	41	49	135
Obesity	44	31	132
Weight loss	43	48	135
Fluid electrolyte	39	47	137
Blood loss anemia	43	51	137
Deficiency anemias	44	41	133
Alcohol abuse	44	35	134
Drug abuse	44	23	133
Psychoses	44	27	132
Depression	44	36	134
Mechanical ventilation	42	44	139
Renal replacement therapy	42	57	138

The final score possible maximum number is 145 (however no patient gets the maximum score in the dataset). The probability of one-year mortality on the validation subset was estimated using the final score as the sole variable in a logistic regression model, model discrimination was examined and the obtained AUROC was 0.785 (95% Confidence Interval: 0.783 - 0.794). To access the calibration of the score, Hosmer–Lemeshow test was used and a value of 0.6 indicating that there is no evidence of poor fit.

The admissions of the validation subset were divided into ten equal size groups according to the increasing estimated probabilities of one-year mortality given by the model, so that in the first group are those admissions that have the lowest probabilities of dying, and in the last group are those admissions with the highest probabilities of dying. For each group the observed and the estimated number of deaths were calculated and compared graphically and presented in Figure 7.14.

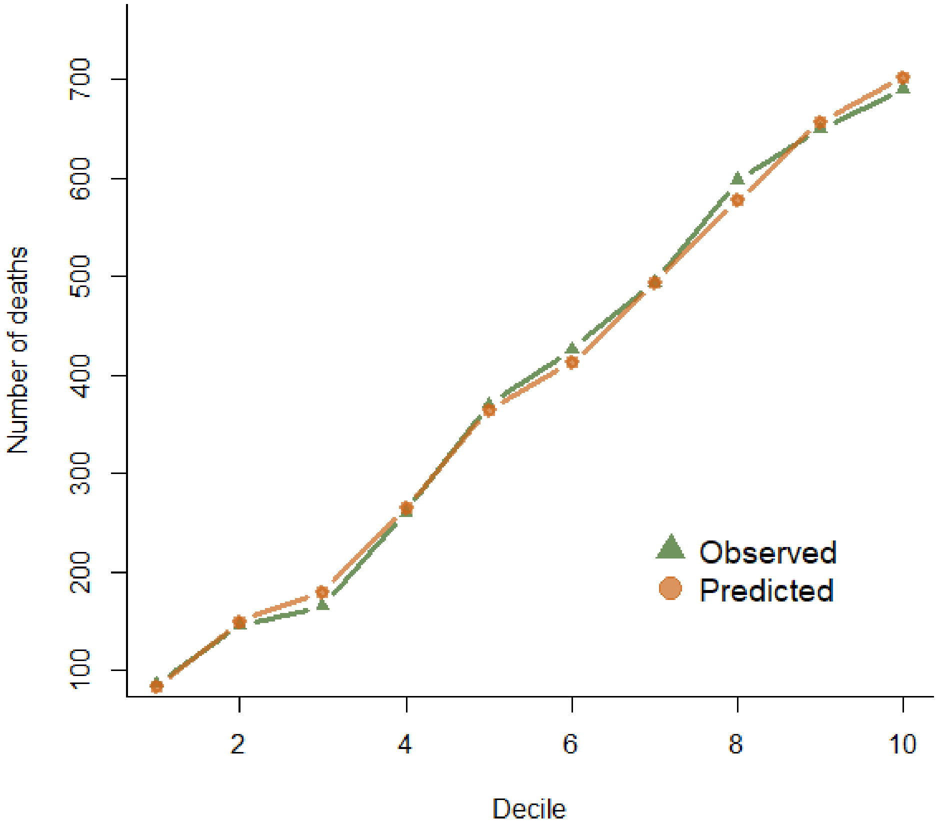


Figure 7.14. Comparison of observed versus predicted number of deaths by groups of increasing probability of one-year mortality.

Some variables, like the temperature or the glucose, have a non-linear behavior with respect to mortality, this approach can detect that kind of behavior and generate multiple CP that allow to interpret the condition of each patient more precisely, in order to prove this, we compared the AUROC the score generated with binary cutoff points and the AUROC of the score generated with multiple cutoff points of over 100 runs, the results are presented in Figure 7.15. In order to evaluate the calibration of the score, we check how the Hosmer-Lemeshow test performs in the said 100 repeated samples, the we calculated the proportion of p-values which are less than 0.05. From 100 runs, the Hosmer-Lemeshow test gave a significant p-value, indicating poor fit, on only 4% of occasions.

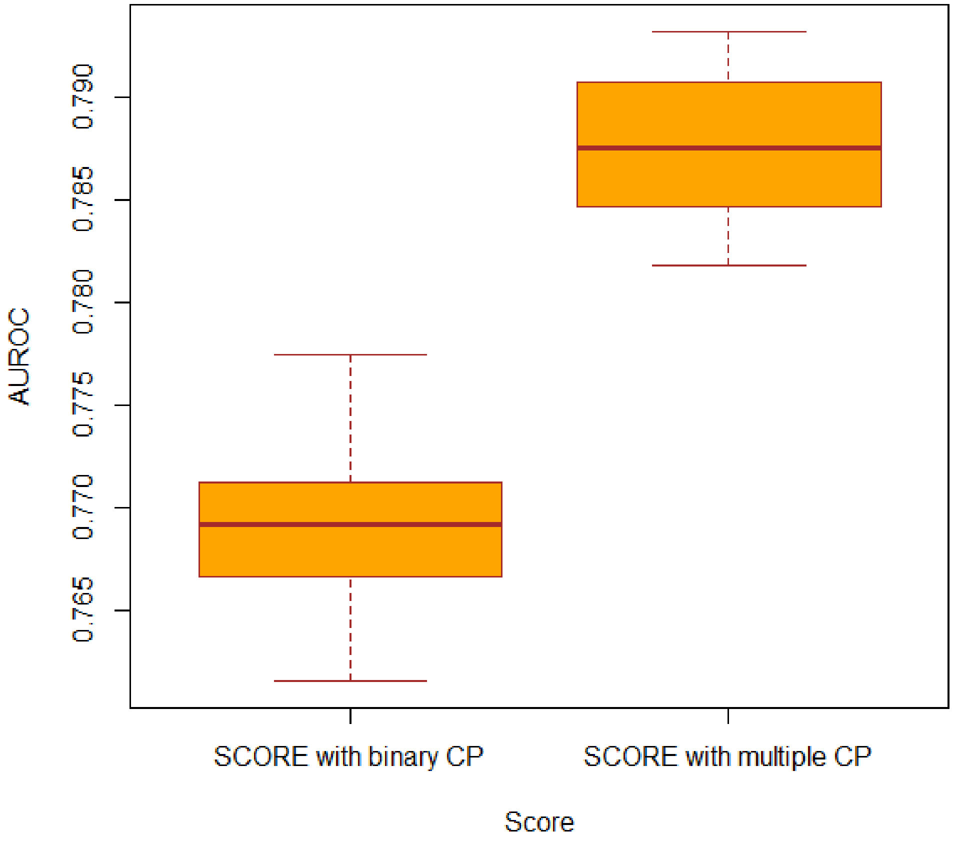


Figure 7.15. AUROC comparison between the score generated with binary cutoff points and the score generated with multiple cutoff point.

7.3 Conclusions

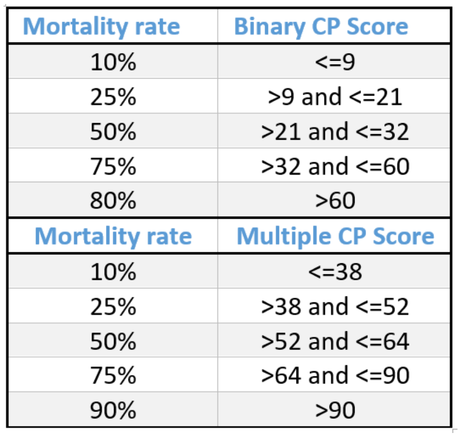
We accessed two severity-of-illness scoring system specifically for patients with sepsis. The scores utilize 6.033 admissions for its development and 9.049 for its validation. The binary cutoff points score contains 52 variables and the multiple cutoff points scores contains 92 variables.

Both scores accurately estimated the probability of one-year mortality in sepsis diagnosed patients within the ICU and are well calibrated. AUROC analysis shows that the presented scores outperforms other scoring systems; even more, the predictive capacity of the score is better, in this study, than the SOFA, that is the scoring systems used for the most recent sepsis and septic shock consensus [17, 18], and the Sepsis Severity Score (SSS), that is an internationally derived scoring system specifically for patients with severe sepsis and septic shock.

The strengths of these scores are that they performed well with respect to both discrimination and calibration. The calibration is especially important as data were collected from a database that spams over 10 years and we use four different sepsis criteria to retrospectively identify the admissions, including the most recent one. The scores are composed of 52 and 92 variables. The number of variables is more than traditional scores like SAPSII (20 variables) and SSS (36 variables) but considerably less than most recent approaches like APACHE IV (142 variables).

The objective of this scores is to early alert of a worse prognostic and to stratify patients according to their risk that can be done according to the Table 7.9. One-year mortality rate according to the developed scores. As expected the multiple CP score presented better discrimination and it use will allow a more accurate interpretation of the condition of each patient, however the binary CP is more easy to implement and can be used for a quick interpretation of a patient's condition.

Table 7.9. One-year mortality rate according to the developed scores.



Mortality rate	Binary CP Score
10%	<=9
25%	>9 and <=21
50%	>21 and <=32
75%	>32 and <=60
80%	>60
Mortality rate	**Multiple CP Score**
10%	<=38
25%	>38 and <=52
50%	>52 and <=64
75%	>64 and <=90
90%	>90

PART 4: PERSONALIZED MODELS

In this part of the thesis we explore the idea that clinical outcome can be personalized and become more precise. Chapter 8 present de development of personalized logistic regression models for this we identify and analyze past patients who have admissions similar to a new patient whose outcome is to be predicted. In chapter 8 we also present an exhaustive evaluation of proposed patient similarity metrics, and concluded that with the correct among of similar patients, determined by a well selected similarity metric, the predictive performance can be improved. Chapter 9 presents the development of personalized stochastic gradient boosting models, that uses the patient similarity metric and the number of patients that proved to perform better in chapter 8. Chapter 10 presents the development of a graph-based regularized multilayer neural network, which is also based on the best performing similarity metric. Chapter 11 present the development of a software based on the characteristics that most contributed to the discrimination of the long-term mortality of patients with sepsis within the ICU.

CHAPTER 8. ONE-YEAR MORTALITY PREDICTION IN PATIENTS ADMITTED TO AN INTENSIVE CARE UNIT WITH DIAGNOSIS OF SEPSIS DRIVEN BY ELECTRONIC MEDICAL DATA AND POPULATION SIMILARITIES

8.1 Introduction

In previous chapters we presented some Severity of Illness Scoring Systems, which are indicators used in the medical practice of an ICU that seek to synthesize information from various physiological and demographic data into a single number that represents the severity of the illness of a patient [2–6]. These indicators are developed from statistical analysis of data collected for a large number of patients; This is the case of the systems for severity of disease classification as SAPS and OASIS, among others. In general, this number increases mortality risk thereof. These classification systems are used to determine the risk in population studies conducted in ICU, and provide a method for benchmarking between intensive care units of different hospitals.

Traditional ICU prediction models are based on the analysis of large populations, and often provide statistically rigorous results for an average patient but are also expensive, time-consuming, and prone to selection bias; moreover, traditionally approaches to ICU outcome prognostication has relied on static models generated from analyzing large, heterogeneous, multi-center patient datasets, such one-size-fits-all approaches perform well for the average patient, but tend to present problems when the characteristics of the patients move away from the average since these indicators lack the precision required for use at the individual level, and they yielded widely dissimilar performances when applied to different groups of patients [62, 115, 116].

In order to mitigate the problems associated with traditional ICU mortality prediction scores, efforts have been made to generate mortality prediction models that use data from patients who share the same characteristic (for example, the same diagnosis or service type) [1, 2, 25, 26]. In the case of sepsis, as mentioned in chapter 2, the performance of mortality prediction systems in patients with suspected severe sepsis and septic shock have been evaluated in the ICU [62], customized versions for severe sepsis and septic shock of in-hospital mortality classification systems have also been developed [62, 63], and even particular scores for the prediction of mortality in patients with severe sepsis and septic shock have been created [64, 65]. In the hospital in general, important studies have been carried out in which exclusive models were developed for the prediction of mortality in patients with sepsis [66–68], for this studies the cohort was not composed exclusively of ICU patients, and although some of the patients received ICU care, the selection criteria are fundamentally different from those of the other studies in which the patients were evaluated for sepsis at the time of admission to the ICU. Although this works report better performance than traditionally severity of disease scores, they focus on the short term mortality prediction (7-day mortality and in-hospital mortality), and the use of in-hospital mortality as an end point for clinical studies are not enough to understand the effect of sepsis on mortality and quality of life [56, 59, 89, 91].

The specific models created according to groups of patients that shares a common sepsis diagnostic have proved to outperform adjusted scoring systems; moreover the models presented in chapter six and seven, have a better predictive capacity with respect to both traditional scoring systems and models created exclusively for patients with sepsis, such as the Severe Sepsis Mortality Prediction Score (SSS) [66]. Besides, even though the SSS scoring system is a severity-of-illness scoring system created specifically for severe sepsis patients it includes for its development both patients with and without ICU stay.

It is clear then, that the presented sepsis mortality prediction models and the ones developed in chapters 6 and 7, continue to be population-based and therefore they provide “the average best choice” for sepsis patients. For this reason, in this chapter we focus on a developing idea in the field of mortality prediction: personalized predictive modeling based on patient similarity. The goal of this approach is to identify patients who are similar to an index patient and derive insights from the data of similar patients to provide personalized predictions. This approach has been widely used for personalized predictions in other fields, including music, movies and e-commerce, however, there are still very few studies that focus on personalized prediction models based on health data prediction.

In a 2017 scoping review, Sharafoddini et al. [117] present the state of techniques in the field of patient similarity in prediction models based on health data. Authors concludes that patient status prediction models based on patient similarity and health data offer exciting potential for personalizing and improving health care, that this field could lead to better patient outcomes and that the interest in patient similarity-based predictive modeling for diagnosis and prognosis has been growing. In contrast, the review includes only 22 articles from 1339 papers that were screened. The selected articles focus on prediction in the health domain, devise a model for prediction, embed explicit patient similarity analytics, and utilize health data for training their model. The dominant focused application areas of the 22 studies reviewed by Sharafoddini et al. are cardiovascular disease (7 studies), diabetes (4 studies), cancer (3 studies) and liver disease (3 studies). The main evaluated outcomes of reviewed articles were diagnosis (9 studies), episode occurrence (4 studies) therapy recommendation (3 studies).

Concretely, in the field of personalized predictive modeling for mortality prediction there was only one reported article. Lee et Al. [2] deployed a cosine-similarity-based patient similarity metric to identify patients that are most similar to an index patient and subsequently custom-build a 30-day mortality prediction model which outperformed the results obtained with models fitted with all the data and traditional severity of disease scores [2]. In their experiments they define 5000 as the minimum number of similar patients for logistic regression to ensure sufficient variability in categorical predictors within training data (these minimum numbers of similar patients could be different for other datasets and predictors) and the best performance (highest AUROC) were achieved with logistic regression when 5000 or 6000 most similar patients were used for training the personalized model. One of the main conclusions of this work is that using a subset of similar patients rather than a larger, heterogeneous population as training data improves mortality prediction performance at the patient level. In this study, predictors equally contribute to the patient similarity metric, the patient cohort is a representation of patients with a wide variety of diagnoses and conditions and a personalized model is fitted for each index admission.

According to the above, in this chapter we present the developing of personalized models that predicts the one-year outcome of sepsis diagnosed patients based on population similarities, moreover, we want to analyze the impact and relevance of the patient similarity metrics when patients are related by a common characteristic (a sepsis diagnosis) and a challenging outcome is evaluated (one-year mortality).

8.2 Methodology

For this study we used Study Cohort B, which mean we used the four criteria to retrospectively identify patients with sepsis within the MIMIC-III database, and the following predictors were included:

- Vital signs : The maximum, minimum and mean values of the following vital signs were extracted during the first 24 hours of the ICU stay: heart rate, systolic blood pressure, diastolic blood pressure, mean blood pressure, respiratory rate, temperature and peripheral capillary oxygen saturation.
- laboratory variables : The maximum and minimum values of following laboratory variables were extracted from the first 24 hours in the ICU: anion gap, bicarbonate, bilirubin, arterial pH, creatinine, chloride, glucose, hematocrit, platelet count, hemoglobin, lactate, potassium, Partial Thromboplastin Time (PTT), Prothrombin Time (PT), the international normalized ratio (INR), sodium, Blood Urea Nitrogen and White Blood Cell Count (wbc).
- Categorical variables: The following categorical variables were extracted: admission type (elective, urgent, emergency), gender, the receipt of either two treatments (dialysis and mechanical ventilation) and comorbidities according to the Elixhauser Comorbidity groups (30 comorbidities).
- Other predictors: The following predictors were also extracted: admission age, the minimum Glasgow Coma Scale, and the total urinary output over the first 24 hours.

Our objective is to use patient similarity to identify a precision cohort for an index admission, which is used to train a personalized model. To do that, we randomly divided the study cohort admissions in a training group with 90% of the admissions and a validation group with the remaining 10%. Each of the admissions in the validation group will be considered as an index admission and its particular precision cohort will be formed by the admissions of the training group that are more similar to said index admission. For each training group, we evaluate five different patient similarity measure in order to select those which the best performance with respect to one-year mortality prediction model.

Each of the admissions of the validation group played the role of the index admissions for which the personalized models were generated. Figure 8.1 depicts the steps executed for processing each admission from the validation group:

1) All pairwise Similarity measures (with five approaches) between the index admission and every admission in the training data were calculated.
2) The calculated similarity values were sorted in ascending order.
3) A precision cohort was created with the data of the n most similar admissions. The number of most similar admission was varied from 1.000, to 13.000 (there are 13.574 admissions in the complete training group

4) Each precision cohort was used to train a logistic regression model for the index admission group.).

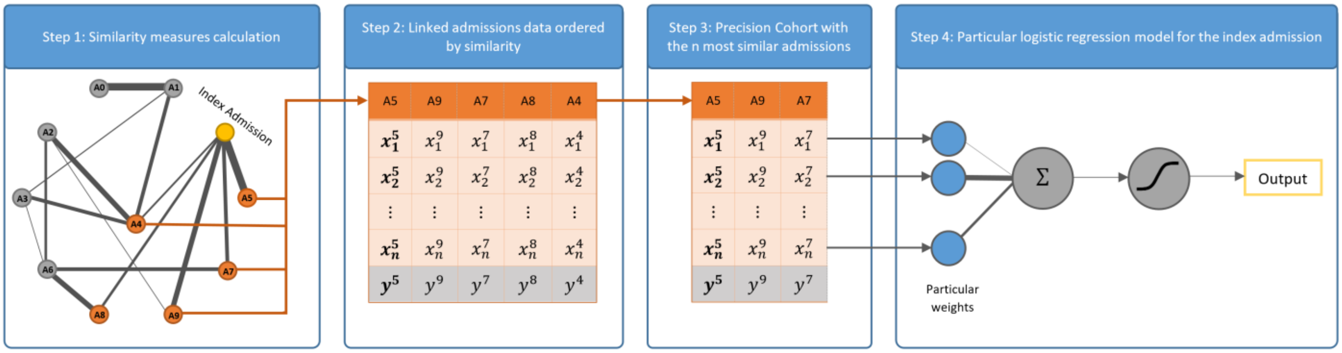

Step 1: Similarity measures calculation
Step 2: Linked admissions data ordered by similarity
Step 3: Precision Cohort with the n most similar admissions
Step 4: Particular logistic regression model for the index admission
Index Admission
A0
A1
A2
A3
A4
A5
A6
A7
A8
A9
A5 A9 A7 A8 A4
x_1^5 x_1^9 x_1^7 x_1^8 x_1^4
x_2^5 x_2^9 x_2^7 x_2^8 x_2^4
x_n^5 x_n^9 x_n^7 x_n^8 x_n^4
y^5 y^9 y^7 y^8 y^4
A5 A9 A7
x_1^5 x_1^9 x_1^7
x_2^5 x_2^9 x_2^7
x_n^5 x_n^9 x_n^7
y^5 y^9 y^7
Σ
Output
Particular weights

Figure 8.1. Overview of the pipeline of the developed model. In Step 1, The index admission is represented by the yellow point; the training admissions are represented by the points labeled A0 to A9; Thickness of the arcs between index-admission node and training- group nodes establish the degree of similarity pairwise. In Step 2 and Step 3, the x_i^j represent the predictors of each linked training admission, and the y^j represents the one-year mortality outcome. In Step 4, the blue circles represent the coefficients for the personalized model.

It is clear that the good performance of this methodology lies in an adequate construction of the precision cohort. For this it is desirable that each index patient has a large number of patients with which to compare (i.e. the training subset to be large), for this reason, we decided to use for our study a training subset formed with 90% of the total admissions of the study B cohort and a test subset with the remaining 10%.

8.2.1 Interaction between admissions

The key aspect of the construction of the precision cohort is the modeling of the interaction between admissions, thus five types of similarity measures were evaluated:

8.2.1.1. Cosine similarity (CS)

Each admission was represented as an Euclidean vector in the multi-dimensional feature space defined by the predictor variables, for this, each continuous predictor was standardized. The similarity between two admissions was defined as follows:

$$Similarity_{cos} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Where A_i and B_i are the components of the vectors of two different admissions, and n is the number of predictors (the extracted clinical and administrative variables). This is equivalent to the similarity used by Lee et al. in [2].

8.2.1.2. Equally Contribution Similarity (ECS)

Since one of the major challenges for population-based studies is comorbidity, and the separation of patients based on demographics and site of care have proved to improve the performance of models [117], we add a similarity term that use a vector composed only by categorical data (comorbidities, treatments, gender and age discretized in age groups) to the CS:

$$Similarity_{eq} = \frac{\sum_{i=1}^{p} A_i B_i}{\sqrt{\sum_{i=1}^{p} A_i^2}\sqrt{\sum_{i=1}^{p} B_i^2}} \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Where A_i and B_i are components of the vectors of two different admissions, n is the number of predictors and p is the number of categorical predictors. The inclusion of such a term achieves that only the admissions that share a common characteristic are connected, moreover, the term reduces or increase the similarity between two patients insofar as they have less or more in common.

8.2.1.3. Weighted Contribution Similarity (WCS)

In ECS all the categorical data equally contribute to the similarity, however, it is clear that different conditions carry different mortality risk; for this reason, a weighted version of the previous approach was also evaluated. Three different set of weights were assessed for the weighted contribution similarity.

$$Similarity_{we} = \frac{\sum_{i=1}^{p} (\theta_i A_i)(\theta_i B_i)}{\sqrt{\sum_{i=1}^{p} (\theta_i A_i)^2}\sqrt{\sum_{i=1}^{p} (\theta_i B_i)^2}} \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Elixhauser Comorbidity Measures

The Elixhauser comorbidity system is a method of categorizing comorbidities of patients based on the International Classification of Diseases (ICD) diagnosis codes found in administrative data, it is composed of 30 comorbidity categories, each of which is dichotomous. Studies have found the Elixhauser comorbidity classification system to be significantly associated with various outcomes including in-hospital mortality [118, 119] and post discharge all-cause mortality [120, 121].

Table 8.1. Elixhauser comorbidity groups and their score association with death in hospital.

Elixhauser Group	Weights(θ)	Elixhauser Group	Weights(θ)
Congestive heart failure	7	**AIDS**	0
Cardiac arrhythmias	5	**Lymphoma**	9
Valvular disease	-1	**Metastatic cancer**	12
Pulmonary circulation	4	**Solid tumor**	4
Peripheral vascular	2	**Rheumatoid arthritis**	0
Hypertension	0	**Coagulopathy**	3
Paralysis	7	**Obesity**	-4
Other neurological	6	**Weight loss**	6
Chronic pulmonary	3	**Fluid electrolyte**	5
Diabetes uncomplicated	0	**Blood loss anemia**	-2
Diabetes complicated	0	**Deficiency anemias**	-2
Hypothyroidism	0	**Alcohol abuse**	0
Renal failure	5	**Drug abuse**	-7
Liver disease	11	**Psychoses**	0
Peptic ulcer	0	**Depression**	-3

In 2009 van Walraven et al. [87] presented a point system for hospital mortality using the Elixhauser comorbidity measures that summarizes all 30 Elixhauser comorbidity groups as a single number to be used for predicting in-hospital mortality. The study included 34.5795 adult admissions between 1996 and 2008 and prove that the Elixhauser comorbidity system can be condensed to a single numeric score that summarizes disease burden and is adequately discriminative for death in hospital. Table 8.1 present the score developed by van Walraven et al. The reported points were used as weights (θ_i) in the weighted contribution similarity metric, the score points that indicate negative association with in-hospital mortality were set as zero, Table 8.1 presents the van Walraven weighted summary score based on the 30 comorbidities from the Elixhauser comorbidity system.

Severe Sepsis Mortality Prediction Score for Use with Administrative Data

In 2016 Ford et al. [66] developed and validated a severe sepsis mortality prediction score using solely administrative data. The score was developed using 563155 admissions based on three criteria for severe sepsis cohort identification (Explicit sepsis, Angus criteria and Martin Criteria). The Sepsis Severity Score (SSS) presented an excellent discrimination for in hospital mortality. Table 8.2 present de variables of the SSS used in the weighted contribution similarity measure.

Table 8.2. Variables of the Sepsis Severity Score used in the weighted contribution similarity measure.

Variable	Weights(θ)	Variable	Weights(θ)
Age < 40 reference	0	**Diabetes complicated**	4
Age>= 40 and <50	8	**Hypothyroidism**	4
Age>= 50 and <60	10	**Renal failure**	6
Age>= 60 and <70	12	**Liver disease**	13
Age>= 70 and <80	15	**Peptic ulcer**	0
Age>= 80 and <90	18	**AIDS**	0
Age>= 90	23	**Lymphoma**	11
Gender: Female	6	**Metastatic cancer**	15
Renal Replacement Therapy	10	**Solid tumor**	10
Mechanical ventilation	23	**Rheumatoid arthritis**	0
Congestive heart failure	5	**Coagulopathy**	0
Cardiac arrhythmias	0	**Obesity**	2
Valvular disease	0	**Weight loss**	3
Pulmonary circulation	6	**Fluid electrolyte**	0
Peripheral vascular	7	**Blood loss anemia**	1
Hypertension	3	**Deficiency anemias**	1
Paralysis	3	**Alcohol abuse**	0
Other neurological	4	**Drug abuse**	1
Chronic pulmonary	4	**Psychoses**	1
Diabetes uncomplicated	4	**Depression**	3

Scoring System for the One-Year Mortality Prediction of Sepsis Patients in Intensive Care Units

In the previous chapter we presented the development of a scoring system for the one-year mortality prediction of sepsis patients in the ICU. The developed score uses the data of 15.082 admissions identified with four sepsis criteria (Explicit sepsis, Angus criteria, Martin Criteria and Sepsis-3) and it outperforms traditional severity of disease scoring systems and even outperform the SSS for the one-year mortality prediction. From the multiple cutoff points score we get the weights that are presented in Table 8.3; for this we subtract 131 to each of the scoring points so the smallest weight was one.

Table 8.3. Weights of the Comorbidities treatments and demographics variables based on the Scoring System for the One-Year Mortality Prediction of Sepsis Patients in ICU.

Variable	Weights(θ)	Variable	Weights(θ)
Admission type: EMERGENCY	10	Lymphoma	10
Gender: Male	7	Metastatic cancer	28
Congestive heart failure	7	Solid tumor	13
Cardiac arrhythmias	6	Rheumatoid arthritis	4
Valvular disease	3	Coagulopathy	4
Pulmonary circulation	5	Obesity	1
Peripheral vascular	7	Weight loss	4
Hypertension	1	Fluid electrolyte	6
Paralysis	4	Blood loss anemia	6
Other neurological	8	Deficiency anemias	2
Chronic pulmonary	5	Alcohol abuse	3
Diabetes uncomplicated	4	Drug abuse	2
Diabetes complicated	3	Psychoses	1
Hypothyroidism	5	Depression	3
Renal failure	6	Mechanical ventilation	8
Liver disease	11	Renal replacement therapy	7
AIDS	6		

8.2.2 Effect of the different similarity measures

We base the studies presented in this chapter on the assumption that, analyzing only similar patients leads to better outcome prediction performance than analyzing all available patients. However, we do not have a ground truth about the similarity between patients, accordingly we evaluate the quality of the proposed patient similarity measures for its ability to generate precision cohorts that lead to the development of models with better predictive capacity, measured by the AUROC; Thus, in this section we present the intuition behind the proposed patient similarity measures but we are only going to determined which of them is the best in future sections in which the performance of the models generated from these metrics is evaluated.

The patient similarity measures yield values between zero (indicating very dissimilar patients) and one (indicating identical patients). When the cosine similarity is used in our dataset, each pair of admissions will have a non-zero number that relates them. In the case of the Equally Contribution Graph and the

Weighted Contribution Similarity each index admission will only be related to those admissions with which it has something in common; for instance, in the Table 8.4 we present the adjacency matrix constructed by using the cosine similarity between an index admission a five training admissions, it is clear that the most similar admission to the index is Adm B, followed by Adm D and the less similar admission is Adm A.

Table 8.4. Cosine similarity example for an index admission.

	Index	Adm A	Adm B	Adm C	Adm D	Adm E
Index	1	0.018	0.252	0.173	0.18	0.023
Adm A	0.018	1	0.13	0.09	0.249	0.135
Adm B	0.252	0.13	1	0.096	0.231	0.244
Adm C	0.173	0.09	0.096	1	0.197	0.311
Adm D	0.18	0.249	0.231	0.197	1	0.403
Adm E	0.023	0.135	0.244	0.311	0.403	1

However, when we analyze the comorbidities, treatments and demographic data, reported in Table 8.5, we find that:

- The index admission and the Adm B only shares the admission type. In similar way, the index admission and the Adm E only shares the admission type.
- The index admission and the Adm D shares hypertension.
- The index admission and the Adm A only have a similar age in common.
- By contrast, the index admission and the Adm C present a similar age, shares the admission type and have seven comorbidities in common: congestive heart failure, cardiac arrhythmias, hypertension, renal failure, liver disease, metastatic cancer and fluid electrolyte.

Table 8.5. Comorbidities, treatments and demographic data of the example admissions. The table only presents the variables in which any of the admissions presents a 1.

Variable	Index	Adm A	Adm B	Adm C	Adm D	Adm E
Admission type: EMERGENCY	1	0	1	1	0	1
Congestive heart failure	1	0	0	1	0	0
Cardiac arrhythmias	1	0	0	1	0	0
Pulmonary circulation	0	0	0	1	0	0
Peripheral vascular	0	1	0	0	0	0
Hypertension	1	0	0	1	1	0
Chronic pulmonary	0	0	0	1	0	0
Diabetes uncomplicated	1	0	0	0	0	0
Diabetes complicated	0	0	0	1	0	0
Renal failure	1	0	0	1	0	0
Liver disease	1	0	0	1	0	0
Metastatic cancer	1	0	0	1	0	0
Coagulopathy	1	0	0	0	0	0
Fluid electrolyte	1	0	0	1	0	0
Deficiency anemias	0	0	0	1	0	0
Alcohol abuse	0	0	0	1	0	0
Drug abuse	0	0	0	0	0	1
Admission age	73.9351	78.9681	24.7912	73.9131	66.0967	21.5041

With this information, we use the second approach - the equally contribution similarity – As result, we could construct a different adjacency matrix, presented in Table 8.6 In this case, as expected, the most similar admission is Adm C.

Table 8.6. Equally contribution similarity example for an index admission.

	Index	Adm A	Adm B	Adm C	Adm D	Adm E
Index	1	0.005	0.084	0.129	0.052	0.006
Adm A	0.005	1	0.037	0.023	0.124	0.061
Adm B	0.084	0.037	1	0.029	0.067	0.189
Adm C	0.129	0.023	0.029	1	0.051	0.072
Adm D	0.052	0.124	0.067	0.051	1	0.18
Adm E	0.006	0.061	0.189	0.072	0.18	1

It can also be interpreted from Table 8.4 and Table 8.6 that the index admission and Adm B presented a similar ICU stay (which means that the laboratory measurements and vital signs presented a similar behavior) but theirs similarity drops when the ECS is applied since they only have the admission type in common; it can also be seen, that the similarities of Adm A and Adm E (that presented a low cosine similarity with the index admission) went to a smaller value.

Index admission and Adm C are expected to remain the most similar when the weighted contribution similarities (WCS) are applied, but the similarity value that represents the relation between the index admission and the other ones should change.

On the other hand, when the Severe Sepsis Mortality Prediction Score (SSS) is used to weight the WCS, the similarity value of Adm B and Adm E are zero because this scoring system do not consider the admission type. Table 8.7 presents the WCS matrix with the SSS weights.

Table 8.7 Weighted contribution similarity example for an index admission with Severe Sepsis Mortality Prediction Score (SSS) weights.

	Index	Adm A	Adm B	Adm C	Adm D	Adm E
Index	1	0.013	0	0.147	0.003	0
Adm A	0.013	1	0	0.032	0.166	0.094
Adm B	0	0	1	0	0	0.01
Adm C	0.147	0.032	0	1	0.003	0
Adm D	0.003	0.166	0	0.003	1	0.344
Adm E	0	0.094	0.01	0	0.344	1

When the Elixhauser Comorbidity Measures (ECM) is used to weight the WCS, the similarity value of Adm A, Adm B, Adm D and Adm E are zero because this weight system do not consider the admission type nor the admission age and the hypertension weight is zero. Since the ECM weight system is the one with fewer variables it also generates the sparest matrix of the evaluated WCS. Table 8.8 presents the WCS matrix with the ECM weights.

Table 8.8 Weighted contribution similarity example for an index admission with Elixhauser Comorbidity Measures (ECM) weights.

	Index	Adm A	Adm B	Adm C	Adm D	Adm E
Index	1	0	0	0.149	0	0
Adm A	0	1	0	0	0	0
Adm B	0	0	1	0	0.254	0.229
Adm C	0.149	0	0	1	0	0
Adm D	0	0	0.254	0	1	0.391
Adm E	0	0	0.229	0	0.391	1

When the Scoring System for the One-Year Mortality Prediction of Sepsis Patients (OMPS) is used, all the training admissions have a non-zero value, since they shared at least one characteristic with the index admission. Table 8.9 presents the WCS matrix with the OMPS weights.

Table 8.9. Weighted contribution similarity example for an index admission with Scoring System for the One-Year Mortality Prediction of Sepsis Patients (OMPS) weights

	Index	Adm A	Adm B	Adm C	Adm D	Adm E
Index	1	0.007	0.08	0.167	0.015	0.006
Adm A	0.007	1	0.026	0.032	0.093	0.053
Adm B	0.08	0.026	1	0.03	0.06	0.202
Adm C	0.167	0.032	0.03	1	0.017	0.081
Adm D	0.015	0.093	0.06	0.017	1	0.202
Adm E	0.006	0.053	0.202	0.081	0.202	1

8.3 Results

A first way to determine which similarity metric results in a better precision cohort for each index admission is to simply use the one-year mortality rate among similar admissions as the prediction. The number of similar admissions were settled as 5, 10, 50, 100, 200, 300,400 and 500.

Figure 8.2 illustrates the AUROC of death counting as a function of the number of similar admissions used as training data, the values presented were deployed using 30 independent runs with different randomly divides portions for training and validation. The maximum mean AUROC of 0.768 (95% confidence interval: 0.744~0.782) was achieved with 100 most similar admissions obtained with the weighted contribution similarity with Scoring System for the One-Year Mortality Prediction of Sepsis Patients weights. The performance degrades rapidly when too few patients are used for training and gradually when more admissions are added to the mortality rate calculation.

The shown trend presented in Figure 8.2 shows that predictive performance based on similarity measures that have the similarity term composed only by categorical data (Equal Contribution similarity and the three versions of the Weighted contribution similarity) is better than when the mortality rate of the 100 most similar patients is used as prediction; however, it seems to flatten when the cosine similarity is used.

To benchmark the personalized one-year mortality prediction models, the predictive performances of ten adjusted severity of disease scoring systems (APSIII, LODS, MLODS, OASIS, qSOFA, SAPS, SAPSII, SIRS, SOFA and SSS) were quantified over 30 independent runs. For each run a different randomly selected training subset (composed of 90% of the total admissions of the study cohort B) were used to adjust the traditional severity of disease scoring systems, and the discrimination performance was obtained over the validation subset (the remaining 10% of admissions); Table 8.10 present this results.

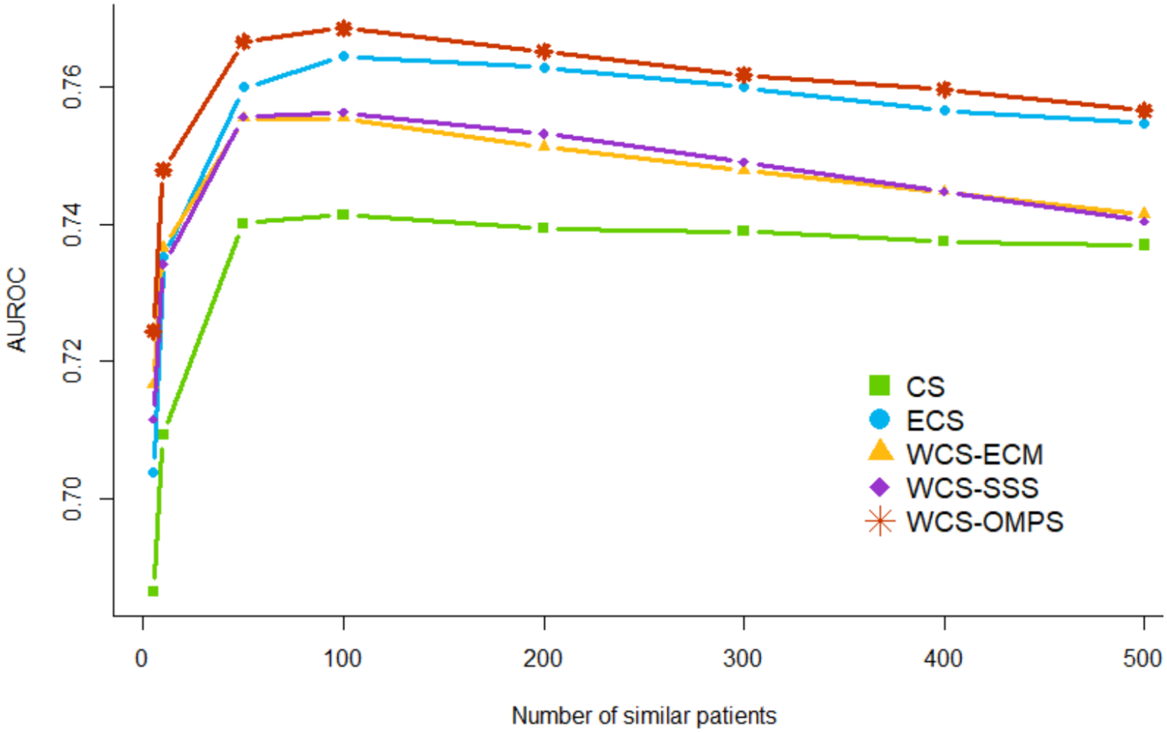


Figure 8.2. Mortality prediction performance of death counting among similar patients. CS: Cosine similarity; ECS: Equal Contribution similarity: WCS-ECM: Weighted contribution similarity with Elixhauser Comorbidity Measures (ECM) weights; WCS-SSS: Weighted contribution similarity with Severe Sepsis Mortality Prediction Score (SSS) weights; WCS-OMPS: Weighted contribution similarity with Scoring System for the One-Year Mortality Prediction of Sepsis Patients (OMPS) weights; AUROC: area under the receiver operating characteristic curve.

Table 8.10. Score performance of different adjusted traditional severity of illness scores for the one-year mortality prediction of sepsis patients within the ICU.

Score	AUROC	95% Confidence Interval
APSIII	0.658	0.631~0.686
LODS	0.631	0.603~0.660
MLODS	0.607	0.578~0.635
OASIS	0.629	0.601~0.658
qSOFA	0.558	0.533~0.584
SAPS	0.626	0.598~0.655
SAPSII	0.700	0.674~0.727
SIRS	0.526	0.498~0.554
SOFA	0.612	0.584~0.641
SSS	0.621	0.593~0.650

The fact that the peak performances presented in Figure 8.2 with all the similarity measures were better than the performance of the adjusted traditional scoring system presented in Table 8.10, indicates that simple death counting among only 100 similar patients resulted in good predictive performance.

The result obtained with the death counting approach corroborate the intuitive idea that similar patients tend to have equal outcomes and proving that the developed patient similarity metrics are adequate for the effective identification of similar patients. However, this approach does not generate a personalized model for the index admission.

In order to generate a personalized one-year mortality model for the sepsis patients within the ICU we used the methodology presented in Figure 8.1. And we executed 30 runs with different randomly selected training (90%) and validation (10%) subsets. On each run, each of the admissions in the validation group was considered as an index admission and a set of precision cohorts were obtained and evaluated. Which this we found the number of similar patients and the similarity measure that presented a better performance.

Figure 8.3 shows the predictive performance of personalized Logistic regression models as a function of the number of similar patients used for training. It is important to highlight that, the AUROC in the last data point (the one on the far right) is equal for all the patient similarity measures, this is because at this point the models are fitted 13574 admissions, what corresponds to the totality of training subset. Moreover, the performance in that data point is equivalent to the performance that would be obtained with the customized multiple cutoff-point score presented in the previous chapter, because such customized scores were developed using logistic regression, and in their development, the performance was conserved regardless of the discretization of the variables.

It can also be observed from Figure 8.3 that that predictive performance improved as a subset of similar patients was used to fit the personalized one-year mortality prediction model. The peak mean AUROC of 0.794 (95% confidence interval: 0.771~0.816) were achieved when 4.000 most similar patients obtained with the weighted contribution similarity with Scoring System for the One-Year Mortality Prediction of Sepsis Patients weights were used to construct the precision cohort.

8.4 Conclusions

The vast among of data that is being stored as Electronic Medical Records enables the development of prediction models based on patient similarities. In this chapter, we presented the utility of similarity metrics in personalizing one-year mortality risk estimation in the ICU for sepsis patients. The results showed that using a subset of similar admission rather than a larger population as training data improves one-year mortality prediction performance, even when the population shares a common characteristic.

Although all the evaluated admissions are from patients with sepsis (which means that they all have an infection and an organ dysfunction), there was improvement when using similarity metrics, even more a simple mortality rate among 100 similar admissions resulted in good predictive performance that exceeded the performance obtained with the scoring systems reported Table 8.10.

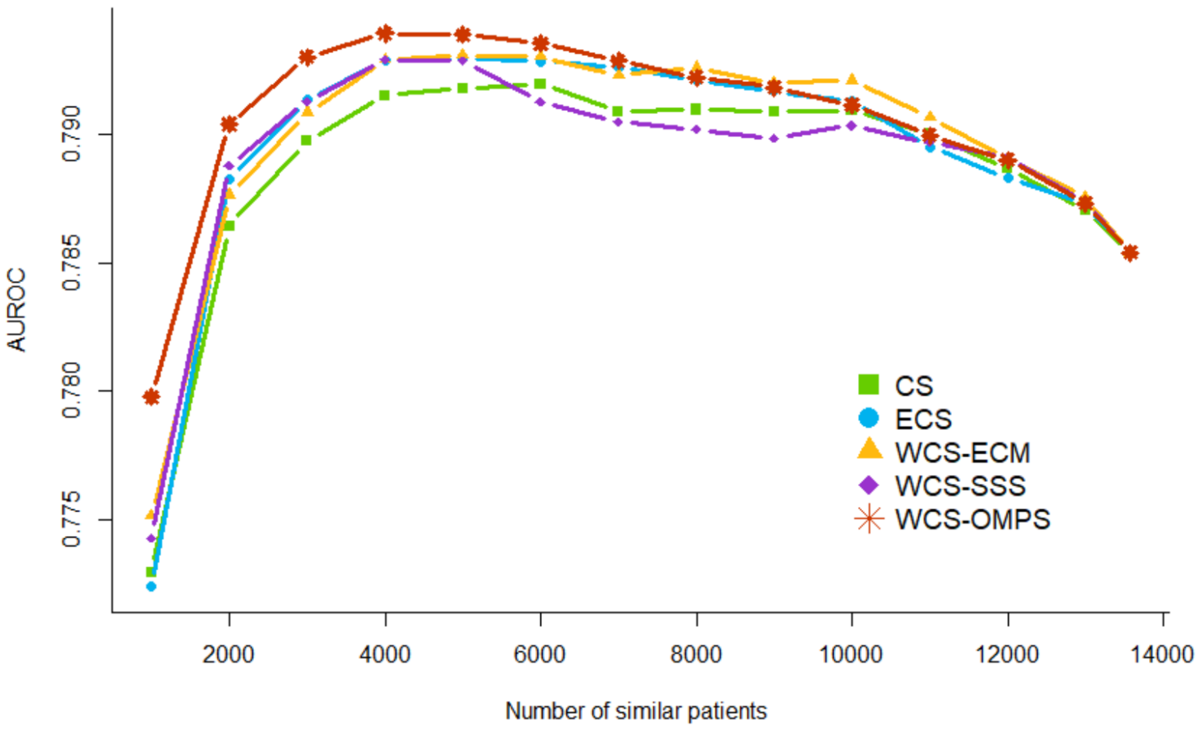


Figure 8.3. One-year Mortality prediction performance of personalized logistic regression trained on similar admissions. CS: Cosine similarity; ECS: Equal Contribution similarity: WCS-ECM: Weighted contribution similarity with Elixhauser Comorbidity Measures (ECM) weights; WCS-SSS: Weighted contribution similarity with Severe Sepsis Mortality Prediction Score (SSS) weights; WCS-OMPS: Weighted contribution similarity with Scoring System for the One-Year Mortality Prediction of Sepsis Patients (OMPS) weights; AUROC: area under the receiver operating characteristic curve.

Traditionally the risk prediction in a ICU, is addressed by the clinician based on large population studies of patients, like severity of disease scores, however the developed models outperformed widely the adjusted traditional used scores, which could be explained by the following elements:

1. The use of a specific cohort of patients with sepsis.
2. The inclusion of sepsis related variables, like lactate.
3. The fact that nearby admissions are more comparable and tend to have the same outcome.

The first two elements explain why the logistic regression model fitted with all the training subset exceeded the performance of the currently used scores, the third one explains the improvements observed in Figure 8.1 and Figure 8.3.

An important aspect of the personalized logistic regression approach is that it gives particular coefficients for each precision cohort which could be interpreted as relative variable importance for a particular patient, meaning that a treating doctor could elucidate the most relevant factor in de prediction, so it has the potential to provide tailored prognoses, and prescribe more effective treatments.

It is clear that one of the factors that strongly affects predictive performance is the choice the similarity measure, the results presented in Figure 8.1 and Figure 8.3 shows that the weighted contribution similarity with Scoring System for the One-Year Mortality Prediction of Sepsis Patients weights generates the personalized one-year mortality prediction models with better predictive performance; however the other weighted contribution similarities did not show a better performance than Equal contribution similarity, one possible explanation could be that the SSS and ECM weights were based on scores that were developed for in-hospital mortality, in addition, the OMPS include all the considered comorbidities and treatments.

Despite the good results reported by Lee et Al. [2], in this study, cosine similarity was, in general, the one that had a worse performance, indicating that importance of the comorbidities when evaluating the long-term mortality. Moreover, the predictive performance improvement reported by Lee et Al. was 2.47% (the best performing model presented an AUROC of 0.83 and the model that used all available data for training presented an AUROC of 0.81), but predictive performance improvement in our study was 1.15%. This could be explained by the fact that the population that we evaluated is especially homogeneous; all of the patients in our study cohort present the same severe diagnosis, sepsis, have a median ICU length of stay of 4 days, a median hospital length of stay of 11 days and a median age of 68 years old.

This study has demonstrated the value of patient similarity-based models in critical health problems and shows the superiority of patient similarity-based models over population-based ones. In order to improve the capabilities of these models we propose as future work, the evaluation of different algorithms on the precision cohort, the implementation of novel machine learning approaches on graph-structured data like graph convolutional networks and the evaluation of different similarity measures.

CHAPTER 9. PERSONALIZED STOCHASTIC GRADIENT BOOSTING MODELS

9.1 Introduction

In previous chapter we presented the developing of personalized models that predicts the one-year outcome of sepsis diagnosed patients based on population similarities, and we concluded that using a subset of similar admission rather than a larger population as training data improves one-year mortality prediction performance, even when the population shares a common characteristic, which means that, despite the fact that the population that we evaluated is homogeneous (Mainly because the data is taken from intensive care unit admissions that share a sepsis diagnosis) the similarity measures are relevant for the long term mortality prediction. We also observed that, from the similarity metrics evaluated, the one that led to the development of models with better performance was the weighted contribution similarity with Scoring System for the One-Year Mortality Prediction of Sepsis Patients weights. In the previous chapter, we also found that the peak AUROC of the personalized logistic regression models were achieved when the data of the 4000 most similar patients were used for training and the 'Weighted Contribution Similarity (WCS)' metric with the 'Scoring System for the One-Year Mortality Prediction of Sepsis Patients (OMPS)' weights were used; proving that clinical long term mortality prediction can become personalized by identifying and training the model with data obtained from past admissions similar to a present case of interest. This idea has also been proven by Lee et al. [122] who suggested that the amount of predictive utility contributed by a past patient should be directly proportional to the degree of similarity between the past and index patient, or if it is seen from other perspective, data from dissimilar patients may actually degrade predictive performance.

In this chapter, we want to evaluate if it is possible to get a more precise long-term outcome prediction using non-linear models supported on the patient similarity measure that proved to enhance the predictive capability of the linear models.

According to the above, in this chapter we present the developing of personalized Stochastic Gradient Boosting models (like the models presented in Chapter 6) that predicts the one-year outcome of sepsis diagnosed patients based on weighted contribution similarity with Scoring System for the One-Year Mortality Prediction of Sepsis Patients weights.

9.2 Methodology

For this study we used Study Cohort B, which mean we used the four criteria to retrospectively identify patients with sepsis within the MIMIC-III database, and the following predictors were included:

- First, the maximum, minimum and mean values of the following vital signs were extracted during the first 24 hours of the ICU stay: heart rate, systolic blood pressure, diastolic blood pressure, mean blood pressure, respiratory rate, temperature and peripheral capillary oxygen saturation.
- Second, the maximum and minimum values of following laboratory variables were extracted from the first 24 hours in the ICU: anion gap, bicarbonate, bilirubin, arterial pH, creatinine, chloride, glucose, hematocrit, platelet count, hemoglobin, lactate, potassium, Partial Thromboplastin Time

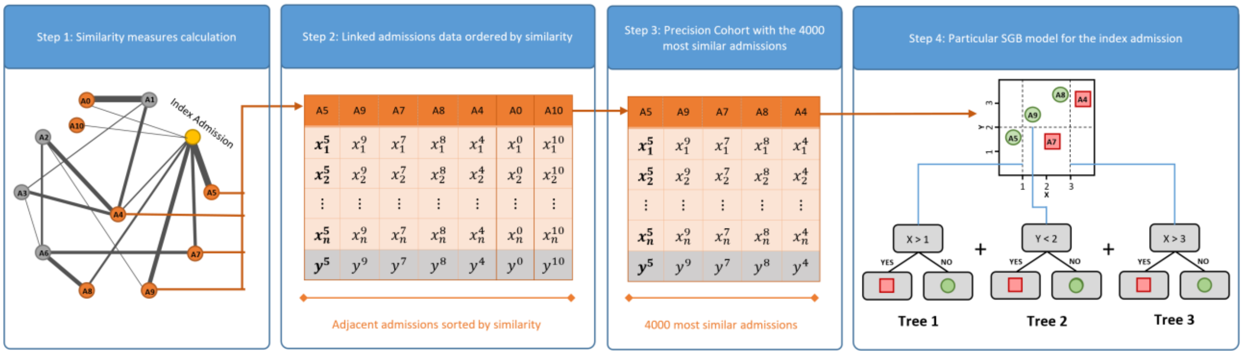
(PTT), Prothrombin Time (PT), the international normalized ratio (INR), sodium, Blood Urea Nitrogen and White Blood Cell Count (wbc).

- Third, the following categorical variables were extracted: admission type (elective, urgent, emergency), gender, the receipt of either two treatments (dialysis and mechanical ventilation) and comorbidities according to the Elixhauser Comorbidity groups (30 comorbidities).
- Lastly, the following predictors were also extracted: admission age, the minimum Glasgow Coma Scale, and the total urinary output over the first 24 hours.

Whit this data we developed personalized stochastic gradient boosting models, for this we used a patient similarity metric to identify a precision cohort for an index admission, which was used to train a personalized non-linear model.

We randomly divided the study cohort admissions in a training group with 90% of the admissions and a validation group with the remaining 10%. For each admission in the validation group, that are considered the index admissions, we executed the following steps (Figure 9.1 presents the overview of this approach):

1. The weighted contribution similarity patient similarity measure with Scoring System for the One-Year Mortality Prediction of Sepsis Patients weights between the index admission and every admission in the training data were calculated.
2. The calculated Similarity values were sorted in ascending order.
3. A precision cohort was created with the data of the 4000 most similar admissions.
4. Each precision cohort was used to train stochastic gradient boosting for the index admission.


Step 1: Similarity measures calculation
Index Admission
A0
A1
A10
A2
A3
A4
A5
A6
A7
A8
A9
Step 2: Linked admissions data ordered by similarity
Adjacent admissions sorted by similarity
Step 3: Precision Cohort with the 4000 most similar admissions
4000 most similar admissions
Step 4: Particular SGB model for the index admission
A8
A4
A9
A5
A7
Y
1
2
3
X
1
2
3
X > 1
YES
NO
+
Y < 2
YES
NO
+
X > 3
YES
NO
Tree 1
Tree 2
Tree 3

A5	A9	A7	A8	A4	A0	A10
x_1^5	x_1^9	x_1^7	x_1^8	x_1^4	x_1^0	x_1^{10}
x_2^5	x_2^9	x_2^7	x_2^8	x_2^4	x_2^0	x_2^{10}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_n^5	x_n^9	x_n^7	x_n^8	x_n^4	x_n^0	x_n^{10}
y^5	y^9	y^7	y^8	y^4	y^0	y^{10}

A5	A9	A7	A8	A4
x_1^5	x_1^9	x_1^7	x_1^8	x_1^4
x_2^5	x_2^9	x_2^7	x_2^8	x_2^4
⋮	⋮	⋮	⋮	⋮
x_n^5	x_n^9	x_n^7	x_n^8	x_n^4
y^5	y^9	y^7	y^8	y^4

Figure 9.1 Overview of the pipeline of the personalized SGB model. In Step 1, The index admission is represented by the yellow point; the training admissions are represented by the points labeled A_0 to A_{10}; Thickness of the arcs between index-admission node and training- group nodes establish the degree of similarity pairwise. In Step 2 and Step 3, the x_i^j represent the predictors of each linked training admission, and the y^j represents the one-year mortality outcome. In Step 4, the green circles represent admissions with a positive outcome and red squares represent admission with a negative outcome, the upper box illustrates a two-dimensional coordinate system, the dotted represent the boundaries of each of the trees in the final model ensemble.

In addition to the technique used for the generation of personalized models in this chapter, an important difference with respect to the methodology followed in Chapter 8, is that both the number of similar patients analyzed for the generation of each model, and the similarity metric used to determine the relationship between patients, remained constant, this was done because in the studies with the personalized logistic regression models we found out that the peak AUROC was achieved when the weighted contribution similarity measure with Scoring System for the One-Year Mortality Prediction of Sepsis Patients weights was used as patient similarity measure, and the 4000 most similar patients were used for training; and since the SGB methodology is computationally more expensive than the logistic regression, it was unfeasible for us to perform tests as rigorous as those performed in the previous chapter; In spite of this, the way in which the tests in this chapter were carried out allow us to conclude whether the inclusion of non-linear methodologies can add value to the prediction of one-year mortality.

The idea behind the selected similarity measure is that different conditions (like comorbidities, treatments or demographics) carry different mortality risk, therefore, it is necessary a metric that allows grouping patients who share conditions according to how related these conditions are to one-year mortality. The following is the used similarity, the weights (θ_i) are presented in Table 9.1.

$$Similarity_{we} = \frac{\sum_{i=1}^{p}(\theta_i A_i)(\theta_i B_i)}{\sqrt{\sum_{i=1}^{p}(\theta_i A_i)^2}\ \sqrt{\sum_{i=1}^{p}(\theta_i B_i)^2}} \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\ \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Table 9.1 Weights of the Comorbidities treatments and demographics variables based on the Scoring System for the One-Year Mortality Prediction of Sepsis Patients in ICU.

Variable	Weights(θ)	Variable	Weights(θ)
Admission type: EMERGENCY	10	**Lymphoma**	10
Gender: Male	7	**Metastatic cancer**	28
Congestive heart failure	7	**Solid tumor**	13
Cardiac arrhythmias	6	**Rheumatoid arthritis**	4
Valvular disease	3	**Coagulopathy**	4
Pulmonary circulation	5	**Obesity**	1
Peripheral vascular	7	**Weight loss**	4
Hypertension	1	**Fluid electrolyte**	6
Paralysis	4	**Blood loss anemia**	6
Other neurological	8	**Deficiency anemias**	2
Chronic pulmonary	5	**Alcohol abuse**	3
Diabetes uncomplicated	4	**Drug abuse**	2
Diabetes complicated	3	**Psychoses**	1
Hypothyroidism	5	**Depression**	3
Renal failure	6	**Mechanical ventilation**	8
Liver disease	11	**Renal replacement therapy**	7
AIDS	6		

The non-linear model used, Stochastic Gradient Boosting (SGB), is the model used for the model presented in chapter 6. Boosting is a powerful machine learning method for selecting features and weight their predictive contribution to the classifier. It combines the outputs of many weak learners, Tress in the case of SGB [103, 123, 124], which are combined through a weighted voting to produce the final prediction. As a decision tree-based ensemble method, boosting allows the use of numeric and categorical predictors and it is robust to missing values [103], it reduces overfitting problems through the use of a learning rate (also called shrinkage) [103], it is also more resistant to multicollinearity than other machine learning methods like neural network [123, 125].

Boosting is based on the idea that the classification error of a model could be reduced if a new weak learner is added. The way of error reduction is arbitrary so that any loss function can be used depending on the problem being solved. In gradient boosting, a gradient descent procedure is used to minimize the loss function. The minimization is realized numerically by applying a steepest descent step that calculating the negative gradient of the loss function [103, 123].

In order to illustrate how the gradient boosting methodology generates the final model, an example of visualizing gradient boosting is presented in Figure 9.2. The points in this example comes from the precision cohort illustrated in Figure 9.1, thus, for the index admission there are five adjacent admissions (3 that survive and 2 that do not), the objective is to generate a model to classify the green circles and red squares, which represent admissions with positive and negative outcomes respectively, shown in the two-dimensional coordinate system. In the first iteration a small tree is generated, this tree divides the two-dimensional space in two segments and indicates that the admissions with a X value less or equal to 1 are classified as green circles. However, with first model two green circles (encased in a red dotted ellipse.) are misclassified. These two green circles are the errors of the first model, for this reason, in the next iteration the algorithm focus on them and generate a second decision tree to correctly predict them; the second tree generates a horizontal boundary in such a way that the admissions that have a Y value higher than 2 are classified as green circles. This second tree separates the three green circles from most of the red squares. However, there is still a misclassified admission, so the third iteration focus on it, and generate a third tree that indicate that the admissions with an X value less or equal to 1 are classified as green circles. The final model is the sum of Tree 1, Tree 2 and Tree 3 and it successfully classifies all the admission.

It is clear then that gradient boosting has the risk of overfitting. This risk can is mitigated by using some regularization methods, the first one is the learning rate υ, which is a number between 0 and 1 that is multiplied to the decision tree generated in each iteration, it has been proven that this parameter improves the model's generalization ability [103]; the second one is the Stochastic Gradient Boosting approach. SGB is a method that applies subsampling as a regulation technique to reduce overfitting [103, 126]. At each iteration SGB samples a fraction of the training data without replacement and uses these subsamples to generate the new tree. Then the improvement of the prediction performance of the new model can be evaluated by predicting those subsamples which are not used in the building of the tree [103, 123]. Besides the learning rate υ, the SGB algorithm involves other parameters which need to be tuned in order to maximize the predictive capability of the model , these parameters are: the total number of boosting iterations (number of trees) and the number of splits performed on each tree [93].

In order to make the generation of personalized SGB models computationally viable, no tuning process was carried out in the development of the particular models of each index admission; instead we selected global hyperparameters and used them in all models. According to the tests performed in chapter 6 we fixed the learning factor as 0.01. To determine and adequate maximum tree depth we generated a SGB model trained with all the training subset and evaluated tree values for the parameter (5,7 and 9). Finally, since it is expected and desirable that a model trained on less admission uses less trees, we developed models using 4000 randomly selected admissions from the training subset which were used to select the final number of trees.

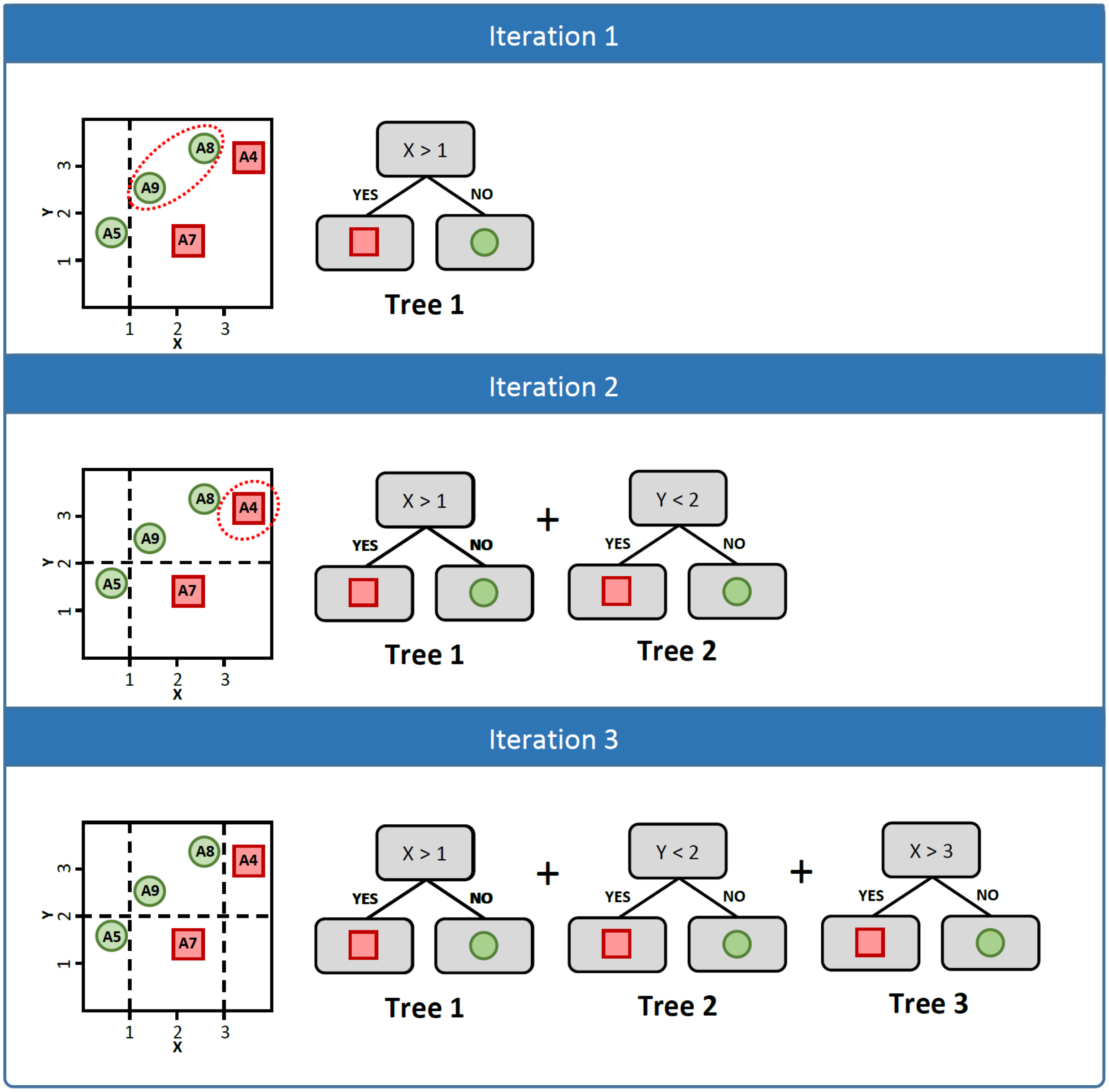


Figure 9.2 Gradient boosting example.

9.3 Results

The development of the personalized stochastic gradient boosting models was done on R software, the caret package [127] was used for the SGB model generation and the proxy package [128] were used to the computation of the similarity measure.

The general SGB model trained with 90% of the admissions included in cohort B reported a AUROC of 0.81 (95% Confidence Interval: 0.78 ~ 0.82) over the remaining 10% of admissions, for this model the learning rate was held constant at a value of 0.01; and a 10 fold cross-validation process was used to select the optimal parameter for the number of splits (tree depth) and boosting iterations, for this the AUROC of different sets of parameter configuration were computed. The results of the performance of 10-fold ross-validation are presented in Figure 9.3.

The parameter that presented a better 10-fold CV performance were: number of trees=1800 and tree depth = 9. However, it can be observed in Figure 9.3 that at 1000 boosting iteration the curve with max tree depth of 9 flattens; moreover, it is reasonable to think that a model trained with less observations (4000 for the personalized SGB models) could need less trees; for this reason, we generate another general SGB model trained with 4000 admissions. This model reported a AUROC of 0.795 (95% Confidence Interval: 0.787 ~ 0.803) over the remaining admissions, for this model the learning rate and tree depth were held constant at 0.01 and 9 respectively; and a 10 fold cross-validation process was used to select the optimal boosting iterations which were set between 600 and 1000 with a stepwise increment of 50. The results of the performance of 10-fold ross-validation are presented in Figure 9.4, where it can be seen that the optimal classifier is constructed with 900 trees.

Figure 9.5 illustrates the AUROC of the personalized models generated with both SGB and logistic regression, the values presented were deployed using 20 independent runs with different randomly divides portions for training and validation. The mean AUROC of the personalized SGB models were 0.809 (95% Confidence Interval: 0.791 - 0.825) and the mean AUROC of the personalized Logistic regression models were 0.794 (95% Confidence Interval: 0.78 - 0.807).

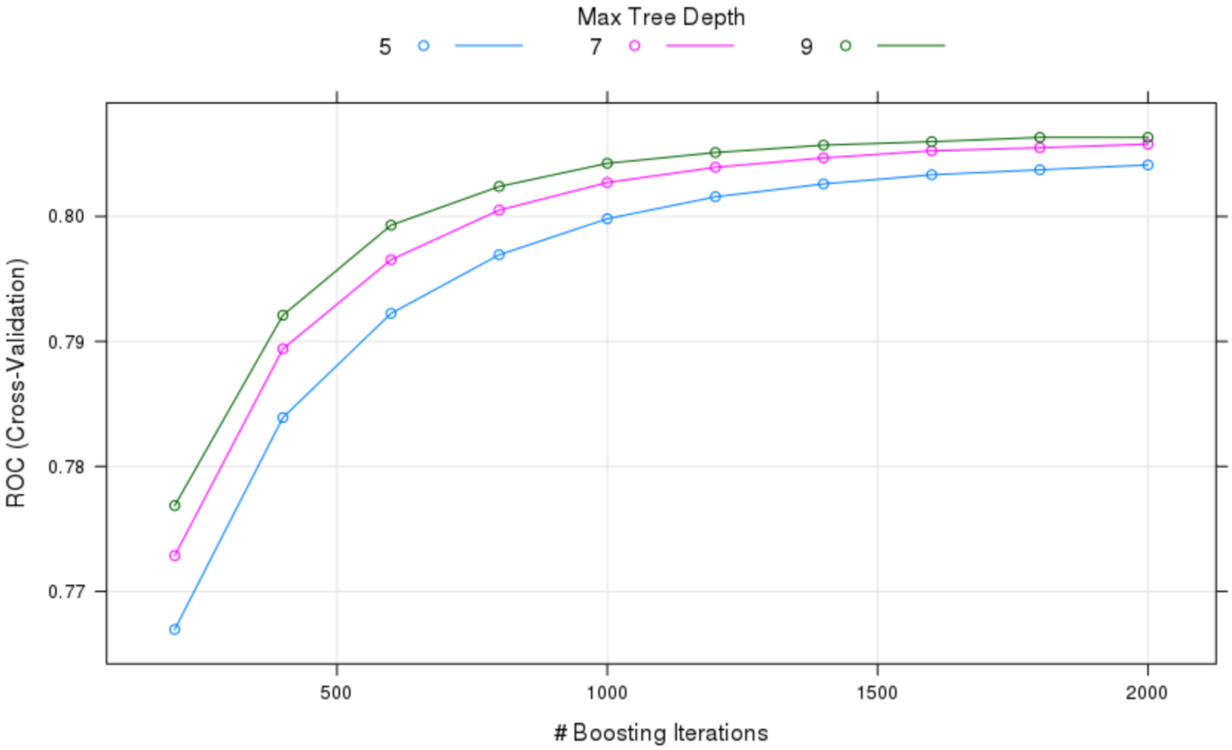


Figure 9.3 SGB model tuning parameters and AUROC. The colored lines indicate different interaction depths (number of splits in each tree). Each data point in the figure represents one evaluated classifier. For instance, the blue data point at (1000, 0.80) indicates a model built with 1000 trees each with 3 splits, that gives a AUROC of 0.80 in the 10-fold cross-validation.

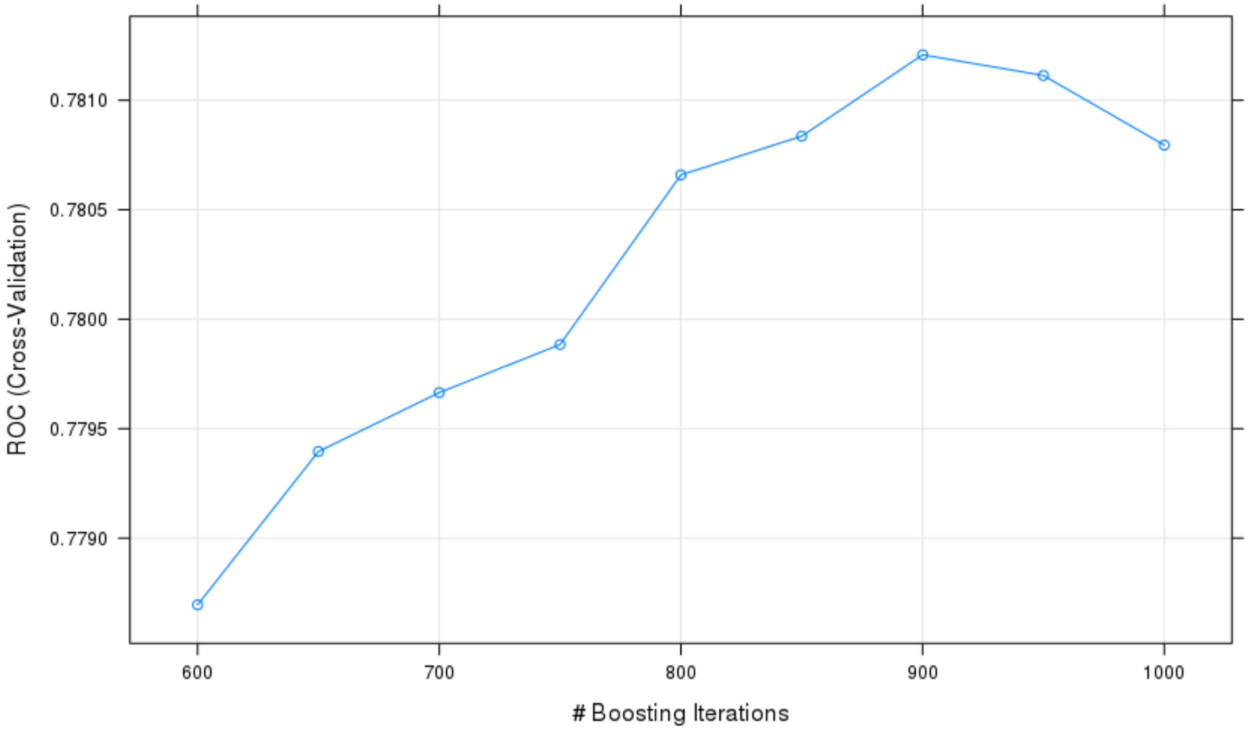


Figure 9.4 SGB model tuning parameters and AUROC.

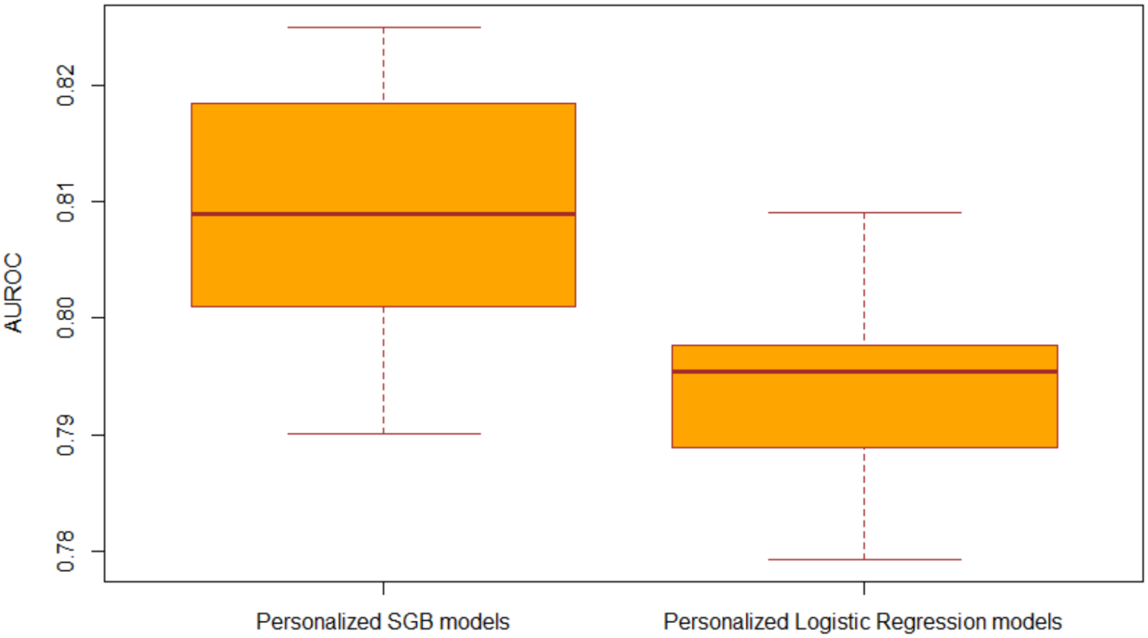


Figure 9.5 AUROC comparison of the personalized model developed with logistic regression and SGB.

9.4 Conclusions

In this chapter we the development of personalized Stochastic Gradient Boosting models. It is important to note that, besides indicating the one-year mortality probability, this approach gives a relative variable importance for each precision cohort, so it has the potential to provide tailored prognoses, and prescribe more effective treatments. As an example we present two particular patients, described in Table 9.2.

Table 9.2 Description of two patients. Only a subset of predictors relevant for the example are presented.

Patient	A	B
One year outcome	Does not survive	Does not survive
Gender	F	M
Cardiac arrhythmias	1	0
Pulmonary circulation	1	0
hypertension	1	0
Diabetes uncomplicated	1	0
Liver disease	0	1
Metastatic cancer	0	0
Solid tumor	1	0
coagulopathy	0	1
Weight loss	1	0
Deficiency anemias	1	0
Alcohol abuse	0	1
Admission age	86.2802	28.6187
Bilirubin maximum	0.9	29.1
Creatinine maximum	1.8	0.8
Hemoglobin maximum	11	13.3
Lactate minimum	2.9	4.9
Lactate maximum	9.6	6.6
Platelet count minimum	272	43
Platelet count maximum	589	123
Partial Thromboplastin Time minimum	28.7	39.2
International Normalized Ratio minimum	1.6	1.9
Blood Urea Nitrogen minimum	35	22
Blood Urea Nitrogen maximum	62	40
Urine Output	1523	705
Heartrate maximum	123	120
Heartrate mean	88.875	104.4328
Systolic blood pressure mean	103.4086	123.0936
Diastolic blood pressure mean	61.60952	63.93333
Mean blood pressure mean	68.78	78.9359
Respiratory rate mean	19.15581	22.3071
Temperature mean	36.5	36.35878
spo2 minimum	90	95
spo2 mean	98.68182	98.52302

Since the personalized SGB model is based on a precision cohort that was constructed specifically for each index patient, it generates two different models, one for the patient A and other for the patient B, and besides the probabilities of surviving or not, it also present additional information.

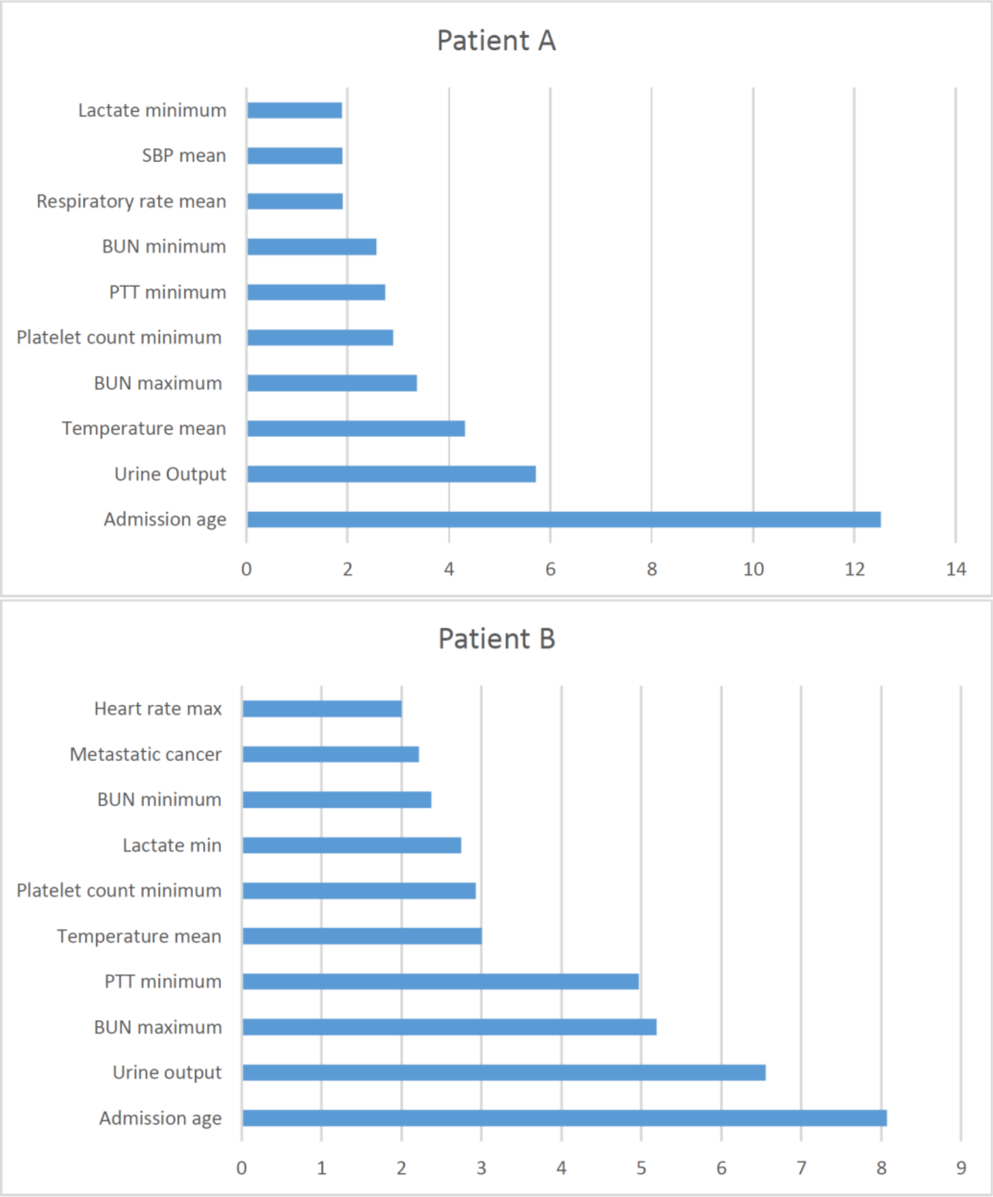


Figure 9.6 Variable importance for the one-year mortality prediction according to the personalized SGB.

Figure 9.6 presents the 10 most important variables for the personalized models of both patients, and differences can be seen bottom up. First the relevance of the admission age is greater in patient A (the older patient). The importance of the blood urea nitrogen is greater in the patient B. There are no variables related with the blood pressure nor the respiratory rate on patient B. And the metastatic cancer appears as a relevant variable in patient B although it does not suffer it, this mean that metastatic cancer is an important predictor in the precision cohort of patient B, which could indicate that this particular comorbidity is one of the main reasons why a young patient can die within a year.

CHAPTER 10. GRAPH-BASED REGULARIZED MULTILAYER NEURAL NETWORK

10.1 Introduction

In this work we are aiming to develop a one-year mortality prediction model for sepsis patients within the ICU that could be used for particular patient prognostication. The approaches that have been implemented can be grouped into three categories: the adjustment of traditional severity of disease scoring systems, the development of entirely new customized models that incorporated additional variables and the generation of personalized models based on a precision cohort for each new patient. Within the approaches used so far, the one that performed best was the personalized Stochastic Gradient Boosting (SGB) models; For the generation of these models, a precision cohort should be constructed, for this, the 4000 patients most similar to each new patient are selected; the similarity between patients is calculated using the weighted contribution patient similarity metric; then, with the constructed precision cohort for the new patient a personalized SGB model is fitted; this means that it is necessary to develop a model for each patient.

According to the above, in this chapter we want to evaluate the possibility of integrating patient similarity information in a model that should only be trained once. For this, we build a structure that contains the vector of characteristics of each patient, and the relationship between patients. Graphs provide a natural way of representing populations and their similarities. In such setting, each patient is represented by a node and the similarities are modelled as weighted edges connecting the nodes [129]. So, our problem is classifying nodes (patients) in a graph; and it can be framed as graph-based learning, where label information is smoothed over the graph using explicit graph based regularization [130].

In this chapter we developed a Graph-based regularized multilayer neural network. For the graph construction we used the same weighted contribution patient similarity metric as in the previous chapter. In order that only the nodes (patients) that are truly similar were connected, we established a similarity threshold, so that the edges that had a value below such threshold would be eliminated from the graph.

10.2 Methodology

For this study we used Study Cohort B, which mean we used the four criteria to retrospectively identify patients with sepsis within the MIMIC-III database, and the following predictors were included:

- First, the maximum, minimum and mean values of the following vital signs were extracted during the first 24 hours of the ICU stay: heart rate, systolic blood pressure, diastolic blood pressure, mean blood pressure, respiratory rate, temperature and peripheral capillary oxygen saturation.
- Second, the maximum and minimum values of following laboratory variables were extracted from the first 24 hours in the ICU: anion gap, bicarbonate, bilirubin, arterial pH, creatinine, chloride, glucose, hematocrit, platelet count, hemoglobin, lactate, potassium, Partial Thromboplastin Time

(PTT), Prothrombin Time (PT), the international normalized ratio (INR), sodium, Blood Urea Nitrogen and White Blood Cell Count (wbc).

- Third, the following categorical variables were extracted: admission type (elective, urgent, emergency), gender, the receipt of either two treatments (dialysis and mechanical ventilation) and comorbidities according to the Elixhauser Comorbidity groups (30 comorbidities).
- Lastly, the following predictors were also extracted: admission age, the minimum Glasgow Coma Scale, and the total urinary output over the first 24 hours.

Similar to the previous chapter we randomly divided the study cohort admissions in a training group with 90% of the admissions and a validation group with the remaining 10%; however, since the methodology developed in this chapter are based on the assumption that similar admissions (nearby nodes in a graph) are more comparable and tend to have the same outcome (labels) in this approach we represent the data of the training subset as a graph.

So, we consider a training subset of 13574 admission comprising demographic, physiological predictors, comorbidities and treatments and we calculated the pairwise similarity between those admissions using the weighted contribution patient similarity metric. The idea behind this patient similarity measure is that different conditions (like comorbidities, treatments or demographics) carry different mortality risk, therefore, it is necessary a metric that allows grouping patients who share conditions according to how related these conditions are to one-year mortality. The following is the used similarity:

$$Similarity_{we} = \frac{\sum_{i=1}^{p}(\theta_i A_i)(\theta_i B_i)}{\sqrt{\sum_{i=1}^{p}(\theta_i A_i)^2}\sqrt{\sum_{i=1}^{p}(\theta_i B_i)^2}} \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Where A_i and B_i are the vectors of two different admissions, n is the number of total predictors and p is the number of categorical predictors (comorbidities, treatments or demographics), and θ_i are the weights of each categorical predictor. In this study we used the scoring system for the one-year mortality prediction of sepsis patients weights, presented in Table 10.1

Then we represent the training subset population as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, where $\mathcal{W}$ is the weighted adjacency matrix describing the graph's connectivity. Each admission in the training subset is represented by a node $v_i \in \mathcal{V}$ and is associated with a feature vector $x(v_i)$ conformed of the predictors (such as laboratory tests, vital signs and comorbidities listed above). The edges $(v_i, v_j) \in \mathcal{E}$ of the graph represent the similarity between the admission. The graph labels are the one-year mortality of the training subset, 1 is for those patients who die before one year and 0 for the patients that survive for more than a year. An overview of the graph generation process is presented in Figure 10.1.

To construct $\mathcal{W}$ we set a similarity threshold (ω_{th}), for this we calculated the pairwise similarity of all the admissions in the training subset; and for each of those admissions we sorted the admissions and found the value of the 4000th most similar patient, and then we average all those values. The objective of this procedure is to adapt the results obtained in chapter 8, where it was shown that the best performing personalized models were achieved when 4000 most similar patients were used to construct the precision cohort for each patient. According to the above $\mathcal{W}$ is defined as:

$$W_{ij} = \begin{cases} Similarity & if \ \ Similarity > \omega_{th} \\ 0 & otherwise \end{cases}$$

Table 10.1 Weights of the Comorbidities treatments and demographics variables based on the Scoring System for the One-Year Mortality Prediction of Sepsis Patients in ICU.

Variable	Weights(θ)	Variable	Weights(θ)
Admission type: EMERGENCY	10	**Lymphoma**	10
Gender: Male	7	**Metastatic cancer**	28
Congestive heart failure	7	**Solid tumor**	13
Cardiac arrhythmias	6	**Rheumatoid arthritis**	4
Valvular disease	3	**Coagulopathy**	4
Pulmonary circulation	5	**Obesity**	1
Peripheral vascular	7	**Weight loss**	4
Hypertension	1	**Fluid electrolyte**	6
Paralysis	4	**Blood loss anemia**	6
Other neurological	8	**Deficiency anemias**	2
Chronic pulmonary	5	**Alcohol abuse**	3
Diabetes uncomplicated	4	**Drug abuse**	2
Diabetes complicated	3	**Psychoses**	1
Hypothyroidism	5	**Depression**	3
Renal failure	6	**Mechanical ventilation**	8
Liver disease	11	**Renal replacement therapy**	7
AIDS	6		

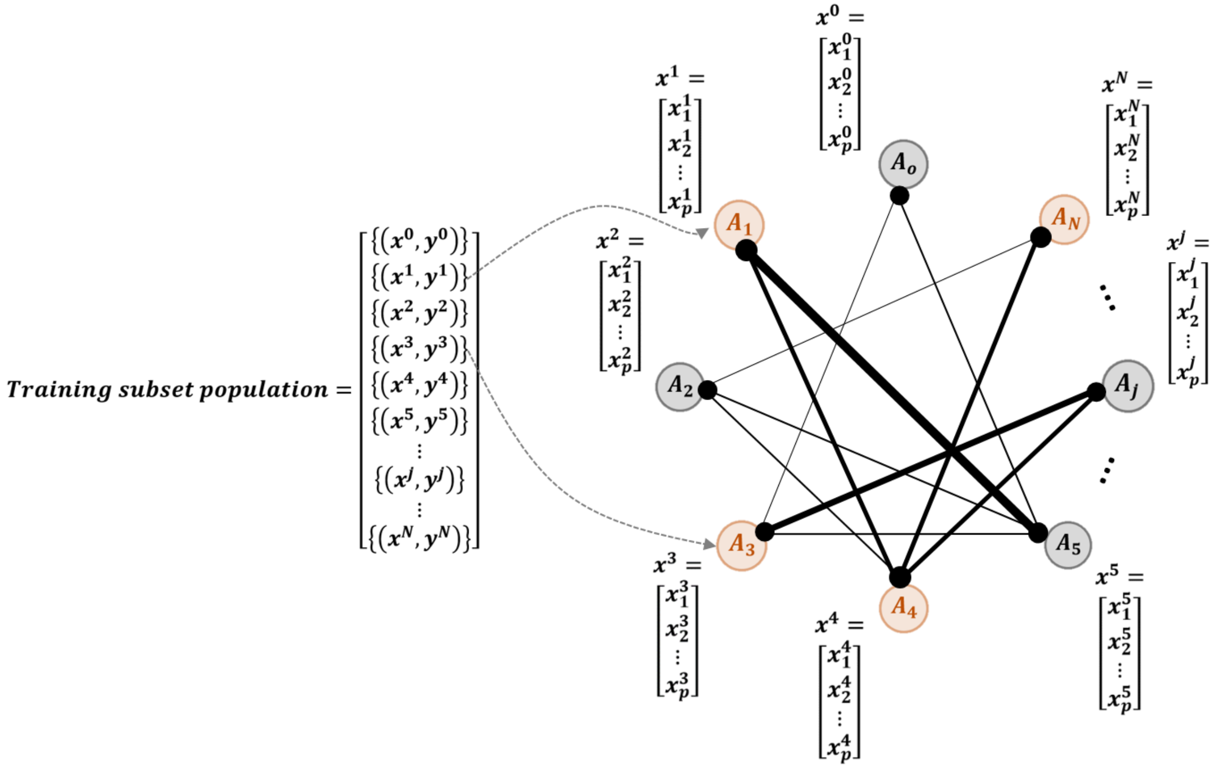
Our problem can be framed as graph-based learning, where label information is smoothed over the graph via some form of explicit graph-based regularization. So it can be addressed as semi-supervised learning a field of study where the goal is to improve generalization (improve the performance) on supervised tasks using unlabeled data, for this, semi-supervised learning algorithms jointly optimize two training objective functions: a supervised loss over labeled data and the unsupervised loss over both labeled and unlabeled data.

There are two semi-supervised learning paradigms, transductive learning and inductive learning. Transductive learning only aims to apply the classifier on the unlabeled instances observed at training time, and the classifier does not generalize to unobserved instances. Inductive learning aims to learn a parameterized classifier that is generalizable to unobserved instances.

In this work we are interested in inductive learning and specifically in those methods that consider that similar instances are more comparable and tend to have the same labels, therefore, our interest is to use, beside the supervised loss term, a loss term that considers the similarity between the admissions. According to the above, we aim to develop a model that incorporates the similarity between patients in the training process, to be used in new patients; meaning that we are not using unlabeled nodes in the training process but we are going to evaluate the performance of the model over a validation subset.

The particular algorithm that we use is called label propagation [131]; this algorithm adds a large penalty when similar instances are predicted to have different labels, the loss function of semi-supervised learning in the binary case can be written as:

$$\sum_{i=1}^{L} l\big(y_i, f(x_i)\big) + \lambda \sum_{i,j} a_{ij} \big\| f(x_i) - f(x_j) \big\|^2$$



$x^0 = [x^0_1, x^0_2, \vdots, x^0_p]$
$x^1 = [x^1_1, x^1_2, \vdots, x^1_p]$
$x^N = [x^N_1, x^N_2, \vdots, x^N_p]$
$x^2 = [x^2_1, x^2_2, \vdots, x^2_p]$
$x^j = [x^j_1, x^j_2, \vdots, x^j_p]$
$x^3 = [x^3_1, x^3_2, \vdots, x^3_p]$
$x^4 = [x^4_1, x^4_2, \vdots, x^4_p]$
$x^5 = [x^5_1, x^5_2, \vdots, x^5_p]$
A_o
A_1
A_N
A_2
A_j
A_3
A_4
A_5
$Training\ subset\ population = [\{(x^0, y^0)\}, \{(x^1, y^1)\}, \{(x^2, y^2)\}, \{(x^3, y^3)\}, \{(x^4, y^4)\}, \{(x^5, y^5)\}, \vdots, \{(x^j, y^j)\}, \vdots, \{(x^N, y^N)\}]$

Figure 10.1 Graph construction process. $\boldsymbol{x^j}$ represent a feature input vector for a particular admission in the training subset and $\boldsymbol{y^j}$ represent the corresponding output. The nodes $\boldsymbol{A_j}$ represent a particular node in the graph with an associated $\boldsymbol{x^j}$ and $\boldsymbol{y^j}$ (which is illustrated as the node color, red nodes indicate a label 1, and grey nodes indicates a label 0).

The first term of the above equation is the standard supervised loss function, for instance the squared loss over the labeled part of the graph, $f(\cdot)$ can be a neural network like differentiable function. The second term is the graph Laplacian regularization, a_{ij} indicates the similarity between instance i and j, so this term incurs a large penalty when similar instances with a large a_{ij} are predicted to have different labels $f(x_i) \neq f(x_j)$.

It is possible to introduce the label propagation idea in a deep learning scheme. Deep learning consists of learning a model with several layers of non-linear mapping. In this chapter we use a multi-layer neural network, and each k^{th} layer is defined as:

$$h_i^k(x) = S\left(\sum_j w_j^{k,i} h_j^{k-1}(x) + b^{k,i}\right), if\ k > 1$$

$$h_i^1(x) = S\left(\sum_j w_j^{1,i} x_j + b^{1,i}\right), if\ k = 1$$

S is a non-linear function such as Re-Lu, w_i^k are the weights associated with each layer, x are the input vector and $b^{k,i}$ are the bias associated with each layer.
The output of the presented neural network for binary classification, assuming N layers of hidden units is a the two position vector:

$$f(x) = \begin{bmatrix} \sum_j w_j^{O,0} h_j^N + b^{O,0} \\ \sum_j w_j^{O,1} h_j^N + b^{O,1} \end{bmatrix}$$

In the implementation of this approach we use a softmax function after the neural network output, which is a function that takes as input a vector of 2 real numbers (for binary classification), and normalizes it into a probability distribution consisting of 2 probabilities. The standard softmax function is defined as:

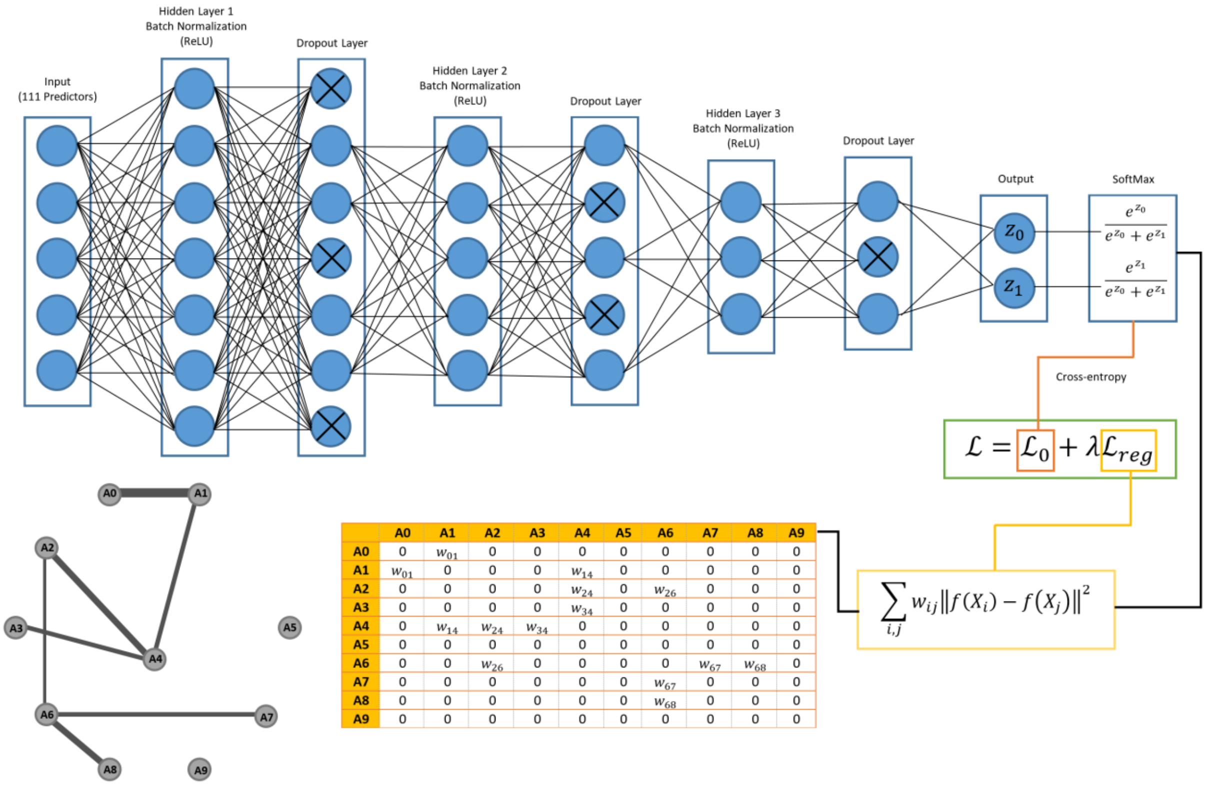
$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}\ for\ i = 1 \ldots K$$

The method we use for deep learning via semi-supervised embedding is to add a semi-supervised loss (regularizer) to the supervised loss on the entire network's output:

$$\mathcal{L} = \mathcal{L}_0 + \lambda \mathcal{L}_{reg},$$
$$\mathcal{L}_{reg} = \sum_{i,j} a_{ij} \left\| f(X_i) - f(X_j) \right\|^2$$

Where $\mathcal{L}_0$is the supervised cross-entropy loss, $f(\cdot)$ is the output of a neural network, λ is a weighing factor and a_{ij} is the similarity weight between the i_{th} admission and the j_{th}admission. Our approach to this problem can be framed as graph-based learning, where label information is smoothed over the graph via explicit graph-based regularization [130, 132]. For this we constructed a graph based on the similarity that presented the best performance in our previous tests. Such graph is denoted as a square matrix A, which has as entries each a_{ij}. Figure 10.2 present the overview of the graph-based regulated neural network. The number of units in the hidden layers were: 20, 15 and 10 respectively and the dropout ratio were 0.5.

For this Laplacian Regularization experiments, we vary the regularization weighting factor λ, which takes the following values:0, 0.0001, 0.001, 0.01, 0.1 and 1;



	A0	A1	A2	A3	A4	A5	A6	A7	A8	A9
A0	0	w_{01}	0	0	0	0	0	0	0	0
A1	w_{01}	0	0	0	w_{14}	0	0	0	0	0
A2	0	0	0	0	w_{24}	0	w_{26}	0	0	0
A3	0	0	0	0	w_{34}	0	0	0	0	0
A4	0	w_{14}	w_{24}	w_{34}	0	0	0	0	0	0
A5	0	0	0	0	0	0	0	0	0	0
A6	0	0	w_{26}	0	0	0	0	w_{67}	w_{68}	0
A7	0	0	0	0	0	0	w_{67}	0	0	0
A8	0	0	0	0	0	0	w_{68}	0	0	0
A9	0	0	0	0	0	0	0	0	0	0

Figure 10.2 Outline of the graph-regulated methodology.

10.3 Results

The first step in for the development of the graph-based regularized multilayer neural network is the construction of the weighted adjacency matrix $\mathcal{W}$. For this we obtain a similarity threshold (ω_{th}) averaging the similarity values of the 4000th most similar patient for each admission in the training subset. The obtained value was 0.08; the distribution of those values is presented in Figure 10.3.

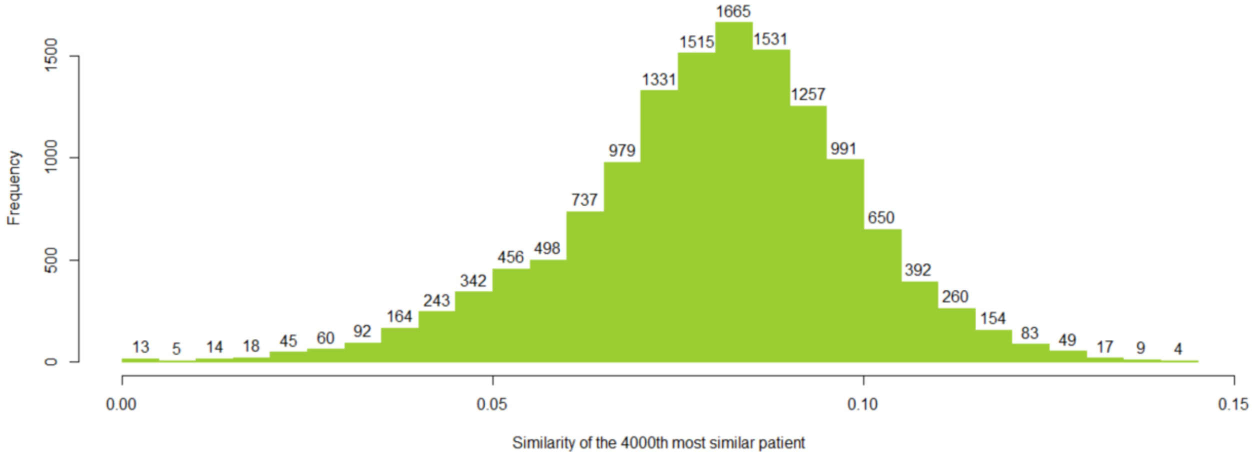


Figure 10.3 distribution of the similarity values of the 4000th most similar patient for each admission in the training subset.

Then, the graph-based regularized multilayer neural network was developed using the TensorFlow [127] library on python, in order to update the network weights iteratively (training) we used the Adam optimization [133] algorithm which is an is an extension to stochastic gradient descent.

Stochastic gradient descent maintains a single learning rate for all weight updates and the learning rate does not change during training, on the contrary, Adam optimization computes individual adaptive learning rates for different parameters.

Unlike the SGB model, where the number of iterations of the algorithm are determinate by the number of trees in the ensemble; in this deep learning approach, the number of iterations were set to 10000000. Each iteration involves using the model with the current weights to make predictions on some samples, comparing the predictions to the ground truth outcomes, calculating the error, and using the error to update the weights by using the backpropagation algorithm. Each iteration was done with a batch of 30, this value was settled by trial and error.

Figure 10.4 presents the performance of the Laplacian regularized neural network as function of the regularization weighting factor λ over ten independent runs. The best performance for the Laplacian regularized neural network was obtained with a regularization factor or 0.1 which reported an AUROC of 0.812 (95% Confidence Interval: 0.808 - 0.814)

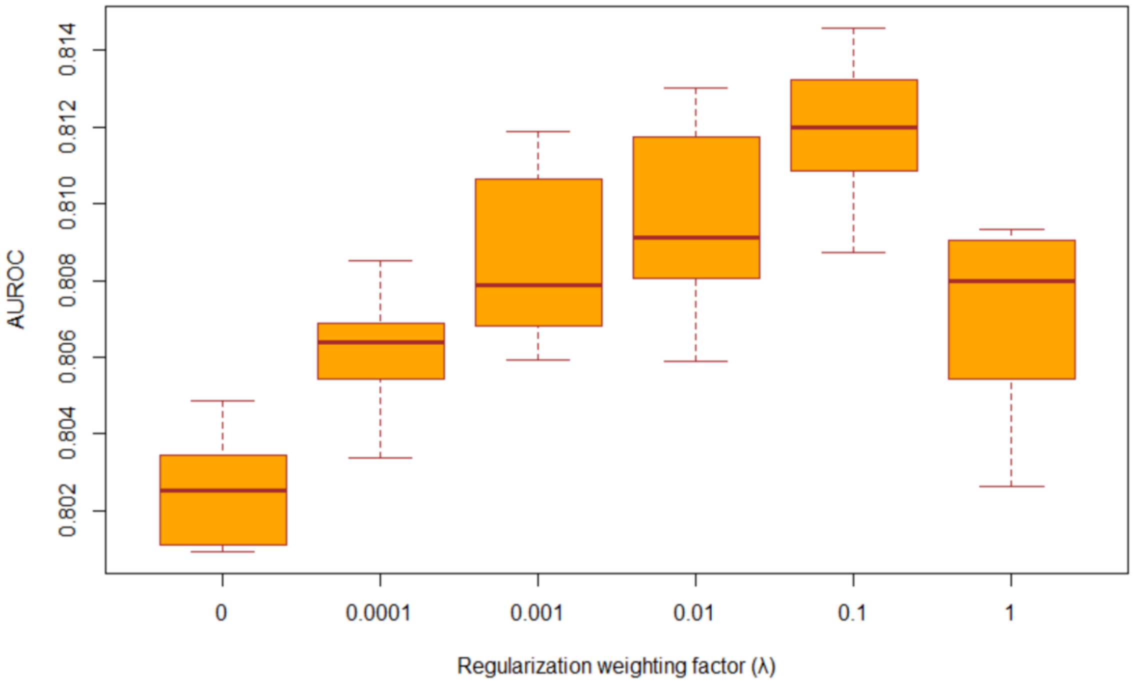


Figure 10.4 Performance of the one-year mortality prediction Laplacian regularized neural network.

10.4 Conclusions

When comparing the graph-based regularized multilayer neural network to the personalized Stochastic Gradient Boosting (SGB) models presented in the previous chapter an important difference arise between them, the personalized SGB models creates a particular model for each new patient, which mean that for every new patient a precision cohort is obtained with the 4000 most similar patients, and a specific SGB model is trained with that particular cohort; and the graph-based regularized multilayer neural network generates a single model that is generalizable to unobserved instances, which mean that it is only trained one time.

When comparing the mean performances of both methodologies it can be observed that the performance obtained with the graph-based regularized multilayer neural network with a weighting factor λ of 0.1 is slightly better than the personalized SGB models. Figure 10.5 illustrates the AUROC of the best performing graph-based regularized multilayer neural network and the personalized models generated with both SGB, it is important to note that the values presented for personalized SGB were deployed using 20 independent runs with different randomly divides portions for training and validation but we only perform 10 runs for the graph-based regularized multilayer neural network with a weighting.

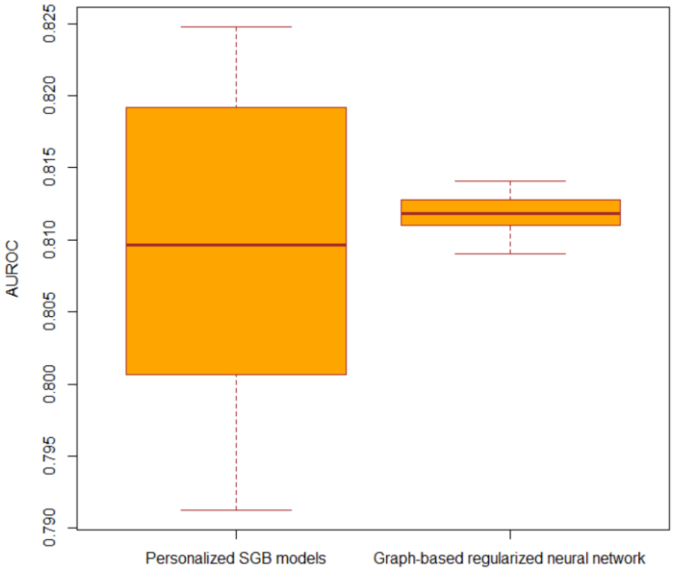


Figure 10.5 Comparison of graph-based regularized multilayer neural network with a weighting factor λ of 0.1 and the SGB personalized models.

The graph-based regularized multilayer neural network has the advantage of being trained only once (in comparison of train a particular model for each index patient), however since it is based on a three-layer neural network loses interpretability. An important factor in favor of the personalized SGB approach is that it could give a relative variable importance for each precision cohort, so it has the potential to provide tailored prognoses, and prescribe more effective treatments.

Despite the fact that the performance of the graph-based regularized multilayer neural network cannot be considered superior to that of personalized SGB, it is clear that it is a promising methodology, and as future work, more rigorous tests are proposed in which other patient similarity metrics are evaluated, the similarity threshold is modified and tests are made with different training subsets.

CHAPTER 11. SOFTWARE DEVELOPMENT

11.1 Introduction

The emergence of machine learning techniques in the field of health is a fact. Specifically, in the field of Intensive Care, it is undeniable that the potential for its application is immense. Specifically, the generation of custom models by groups of populations, the use of assembly algorithms and the implementation of patient similarity based models, contribute elements to the prediction of mortality that lead to a better identification of patients at risk and therefore to improve health services.

In previous chapters we conclude that the generation of one-year mortality prediction scores exclusively for sepsis patients within the ICU widely outperforms adjusted traditional severity of illness scoring systems [7, 8]. In addition, we demonstrated that the use of well selected similarity measure for the generation of personalized models improves the discrimination performance moreover, the use of non-linear models, like SGB, could indicate which variables are more important for each personalized model.

In this chapter we present a software development based on the characteristics that most contributed to the discrimination of the long-term mortality of patients with sepsis within the ICU, and provide information of clinical utility, according to the criterion of intensivist experts. The goal of this software is to enable the use of the developed models in a clinical environment and presents three personalized outputs for each new patient: the one-year mortality rate among the 100 most similar patients, an estimate of the one-year mortality probability based on a well selected precision cohort, and the 10 most relevant parameters for the precision cohort. These outcomes are based on the similarity measures presented in previous chapters, and the complete data of the cohort B.

11.2 Software functionality

According to the results presented in previous chapters, we identify three features to include in the software:

- Personalized model: the software generates a personalized model that predicts the one-year outcome of sepsis-diagnosed patients. it generates a precision cohort for the new patient, the precision cohort is constituted by the 4000 most similar admissions; the similarity between the new patient and the 15082 admissions available in the database are computed using weighted contribution similarity with Scoring System for the One-Year Mortality Prediction of Sepsis Patients weights.
- Relative variable importance: Each precision cohort was used to train a particular Stochastic Gradient Boosting (SGB) [123, 134] model for the new patients. Although the resulting SGB model it is complex, since it is composed of 900 trees, SGB model has good interpretability, being as, it can identify the variable importance, and because the model is generated from the precision cohort for each new patient, the model would provide the most important predictors for each particular case.

- One-year mortality rate among the 100 most similar patients: another way to determine how is the patient's condition with respect to his closest neighbors is using the mortality rate, for this we simply use the one-year mortality rate among the 100 most similar admissions, a method that, when used as prediction for the new patient, proved to outperform adjusted traditional severity of illness scoring systems.

The developed software is an interactive web application, constructed using the Shiny R package [135]. And it dynamic is divided in three layers. The first one, the presentation layer, is with which the user interacts directly and consist of the data input and visualization modules. The second one is the control layer, in this one the precision cohort is obtained and the models are generated. The final one is the data layer, and it is where the database composed of all sepsis admissions of cohort B is located. Figure 11.1 shows the software dynamics, and the interaction between layers.

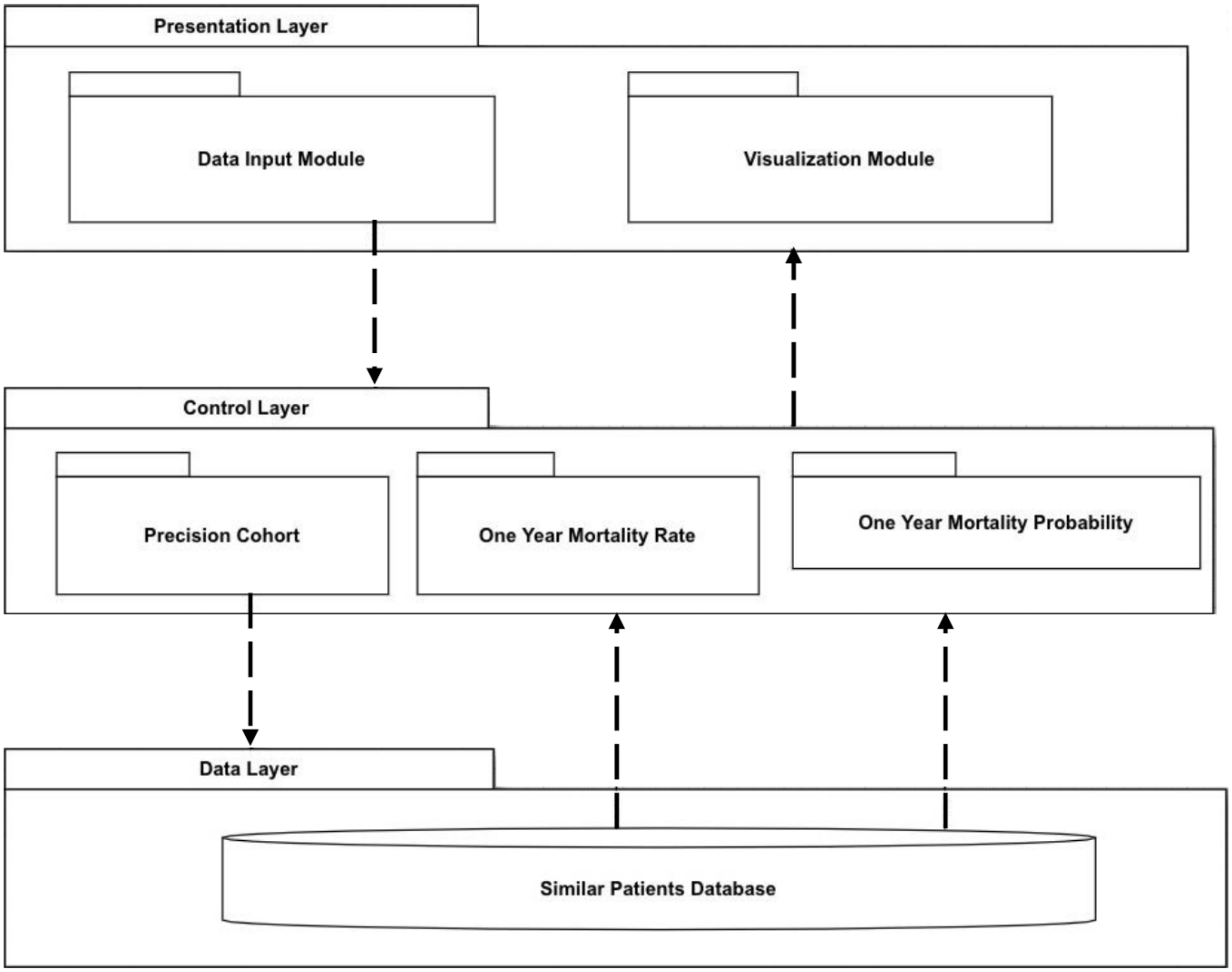


Figure 11.1. Dynamics of the software

The interface of the software is composed of two main panels, one for the input of the variables, and another one for the results display. The software has four modules. The first one enable user to insert the new patient data, which are assigned to a feature vector, the input data are divided in four categories:

- Admission data and demographics: categorical variables as admission type (elective, emergency), gender, birthdate.
- Comorbidities and treatments: Two treatments are included as input variables (dialysis and mechanical ventilation) and the comorbidities according to the Elixhauser comorbidity groups [87] (30 comorbidities).
- Routine charted data: the maximum, minimum and mean values of the following vital signs were extracted during the first 24 hours of the ICU stay: heart rate, systolic blood pressure, diastolic blood pressure, mean blood pressure, respiratory rate, temperature and peripheral capillary oxygen saturation. Also, the total urine output during the first 24 hours and the minimum Glasgow Coma Scale Score during the first 24 hours were included as input variables.
- Laboratory based measures: the maximum and minimum values of following laboratory variables were extracted from the first 24 hours in the ICU: anion gap, bicarbonate, bilirubin, arterial pH, creatinine, chloride, glucose, hematocrit, platelet count, hemoglobin, lactate, potassium, Partial Thromboplastin Time (PTT), Prothrombin Time (PT), the international normalized ratio (INR), sodium, Blood Urea Nitrogen and White Blood Cell Count (wbc).

The second module selects the precision cohort; In this case, the similarity between the new patients and all the 15082 admissions of the database are computed obtaining a similarity vector that is sorted. Then, we use the sorted indices to obtain a similarity-sorted database, when the most similar patients come first. For the similarity computation we used the Weighted Contribution Similarity, a similarity measure in which each categorical data (like the comorbidities, treatments and demographics) had a particular weight, contributing more or less to the similarity:

$$Similarity_{we} = \frac{\sum_{i=1}^{p}(\theta_i A_i)(\theta_i B_i)}{\sqrt{\sum_{i=1}^{p}(\theta_i A_i)^2}\sqrt{\sum_{i=1}^{p}(\theta_i B_i)^2}} * \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

The weights used for the categorical data are based on the Scoring System for the One-Year Mortality Prediction of Sepsis Patients in Intensive Care Units and presented in Table 11.1.

The third module computes the one-year mortality rate among the 100 most similar patients; for this we took the one-year outcome of the top 100 patients in the similarity sorted database (which are a zero for surviving patients and one for non-surviving patients) sum the all and multiply the result by 100, this is presented in the software as a pie chart.

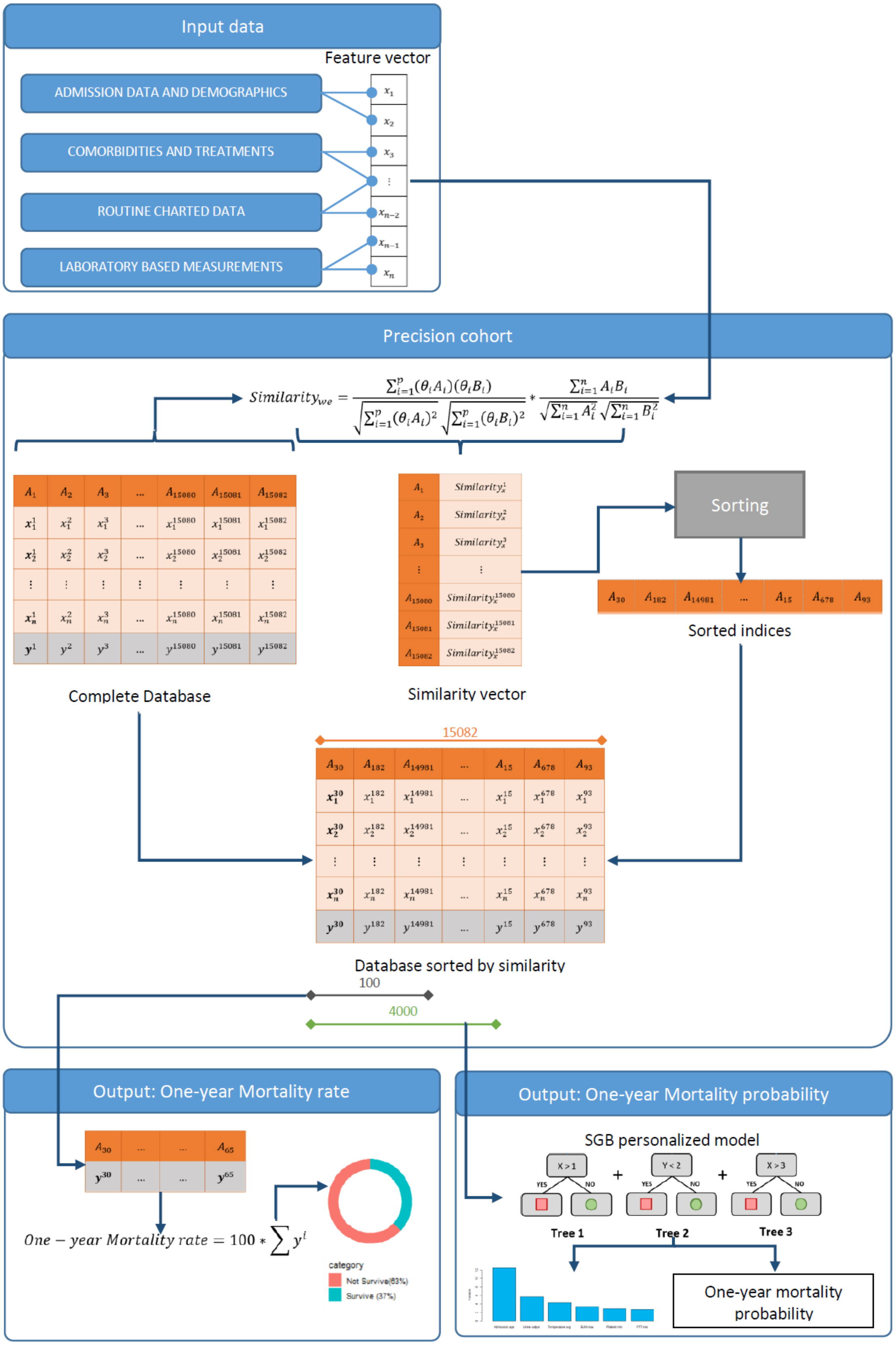
The last module computes the one-year mortality probability; for this a personalized SGB model is fitted with the precision cohort (the data of the 4000 most similar patients). Along the mortality probability we also present the relative variable importance of the top 10 most relevant predictor for each particular precision cohort.

Table 11.1. Weights of the Comorbidities treatments and demographics variables based on the Scoring System for the One-Year Mortality Prediction of Sepsis Patients in ICU.

Variable	Weights(θ)	Variable	Weights(θ)
Admission type: EMERGENCY	10	**Lymphoma**	10
Gender: Male	7	**Metastatic cancer**	28
Congestive heart failure	7	**Solid tumor**	13
Cardiac arrhythmias	6	**Rheumatoid arthritis**	4
Valvular disease	3	**Coagulopathy**	4
Pulmonary circulation	5	**Obesity**	1
Peripheral vascular	7	**Weight loss**	4
Hypertension	1	**Fluid electrolyte**	6
Paralysis	4	**Blood loss anemia**	6
Other neurological	8	**Deficiency anemias**	2
Chronic pulmonary	5	**Alcohol abuse**	3
Diabetes uncomplicated	4	**Drug abuse**	2
Diabetes complicated	3	**Psychoses**	1
Hypothyroidism	5	**Depression**	3
Renal failure	6	**Mechanical ventilation**	8
Liver disease	11	**Renal replacement therapy**	7
AIDS	6		

Figure 11.2 depicts modules that compose the software; in the input data module, the four collapsible panels and the feature are presented; in the precision cohort module, the first matrix is the complete dataset, where the dark orange row represent the indices of each admission, the light orange rows represent the predictors and the grey row represents the one-year mortality outcomes of each admission. In the similarity vector the dark orange column represent the indices of the admissions and the light orange column represents the similarity between each admission and the new patient; from the sorting process we only use the indices to rearrange the complete dataset and obtain the database sorted by similarity.

The one-year mortality rate outcome module uses a vector with the one-year mortality outcomes of the 100 most similar admissions, i.e. the first one hundred columns of the database sorted by similarity. Finally, the one-year mortality probability outcome uses a precision cohort composed of the data of the 4000 most similar admissions to the new patient and fit a personalized SGB model, which is represented by a sum of trees, and return two outputs: the one-year mortality probability, and the relative importance of all the predictors, the top ten most important predictors are displayed in a bar plot.



$$Similarity_{we} = \frac{\sum_{i=1}^{p}(\theta_i A_i)(\theta_i B_i)}{\sqrt{\sum_{i=1}^{p}(\theta_i A_i)^2}\sqrt{\sum_{i=1}^{p}(\theta_i B_i)^2}} * \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

A_1	A_2	A_3	...	A_{15080}	A_{15081}	A_{15082}
x_1^1	x_1^2	x_1^3	...	x_1^{15080}	x_1^{15081}	x_1^{15082}
x_2^1	x_2^2	x_2^3	...	x_2^{15080}	x_2^{15081}	x_2^{15082}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_n^1	x_n^2	x_n^3	...	x_n^{15080}	x_n^{15081}	x_n^{15082}
y^1	y^2	y^3	...	y^{15080}	y^{15081}	y^{15082}

Complete Database

A_1	$Similarity_x^1$
A_2	$Similarity_x^2$
A_3	$Similarity_x^3$
⋮	⋮
A_{15080}	$Similarity_x^{15080}$
A_{15081}	$Similarity_x^{15081}$
A_{15082}	$Similarity_x^{15082}$

Similarity vector

A_{30}	A_{182}	A_{14981}	...	A_{15}	A_{678}	A_{93}

Sorted indices

A_{30}	A_{182}	A_{14981}	...	A_{15}	A_{678}	A_{93}
x_1^{30}	x_1^{182}	x_1^{14981}	...	x_1^{15}	x_1^{678}	x_1^{93}
x_2^{30}	x_2^{182}	x_2^{14981}	...	x_2^{15}	x_2^{678}	x_2^{93}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_n^{30}	x_n^{182}	x_n^{14981}	...	x_n^{15}	x_n^{678}	x_n^{93}
y^{30}	y^{182}	y^{14981}	...	y^{15}	y^{678}	y^{93}

Database sorted by similarity

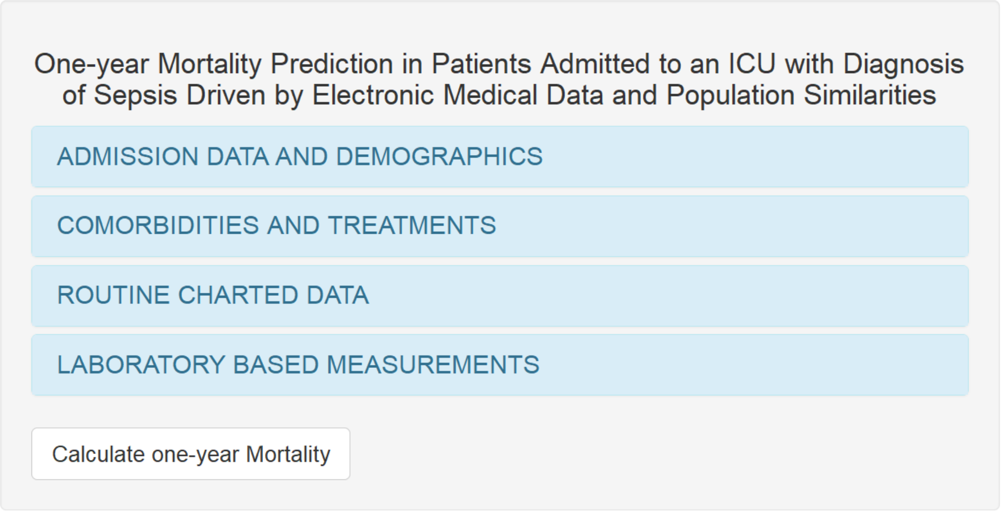
A_{30}	...	...	A_{65}
y^{30}	...	...	y^{65}

$$One-year\ Mortality\ rate = 100 * \sum y^i$$

Figure 11.2. structure of the software developed.

11.3 Results

As mentioned before the software is a web application composed of two panels the first one is used for the input of the data, and it is composed of 4 collapsible panels that correspond to each of the four variable categories. Figure 11.3 present the final appearance of said panel.

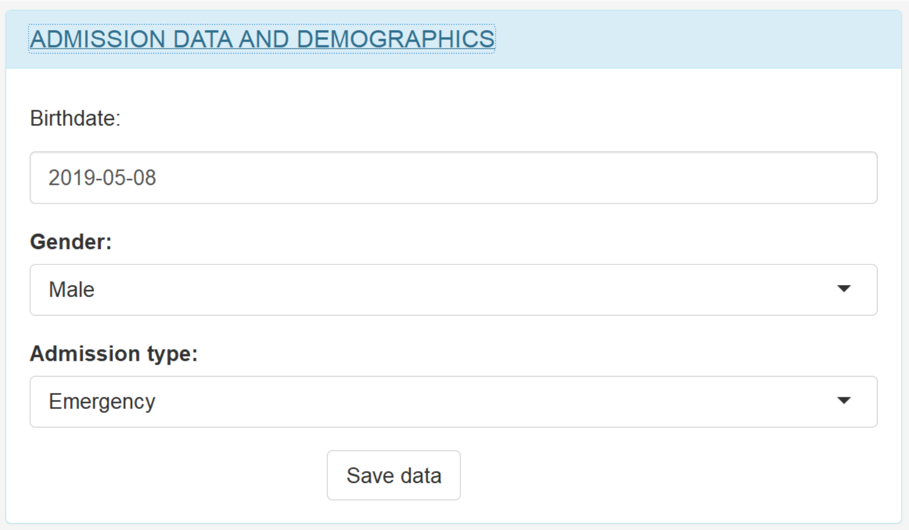


One-year Mortality Prediction in Patients Admitted to an ICU with Diagnosis of Sepsis Driven by Electronic Medical Data and Population Similarities

ADMISSION DATA AND DEMOGRAPHICS
COMORBIDITIES AND TREATMENTS
ROUTINE CHARTED DATA
LABORATORY BASED MEASUREMENTS
Calculate one-year Mortality

Figure 11.3. Input panel.

The admission data and demographics collapsible panel is composed of a date input element, where the user enters the patient birthdate, and it is transformed to the admission age. It also contains two selection inputs, one for the gender (Male or Female) and one for the admission type (Emergency, Elective). Figure 11.4 presents the final appearance of the admission data and demographics collapsible panel.



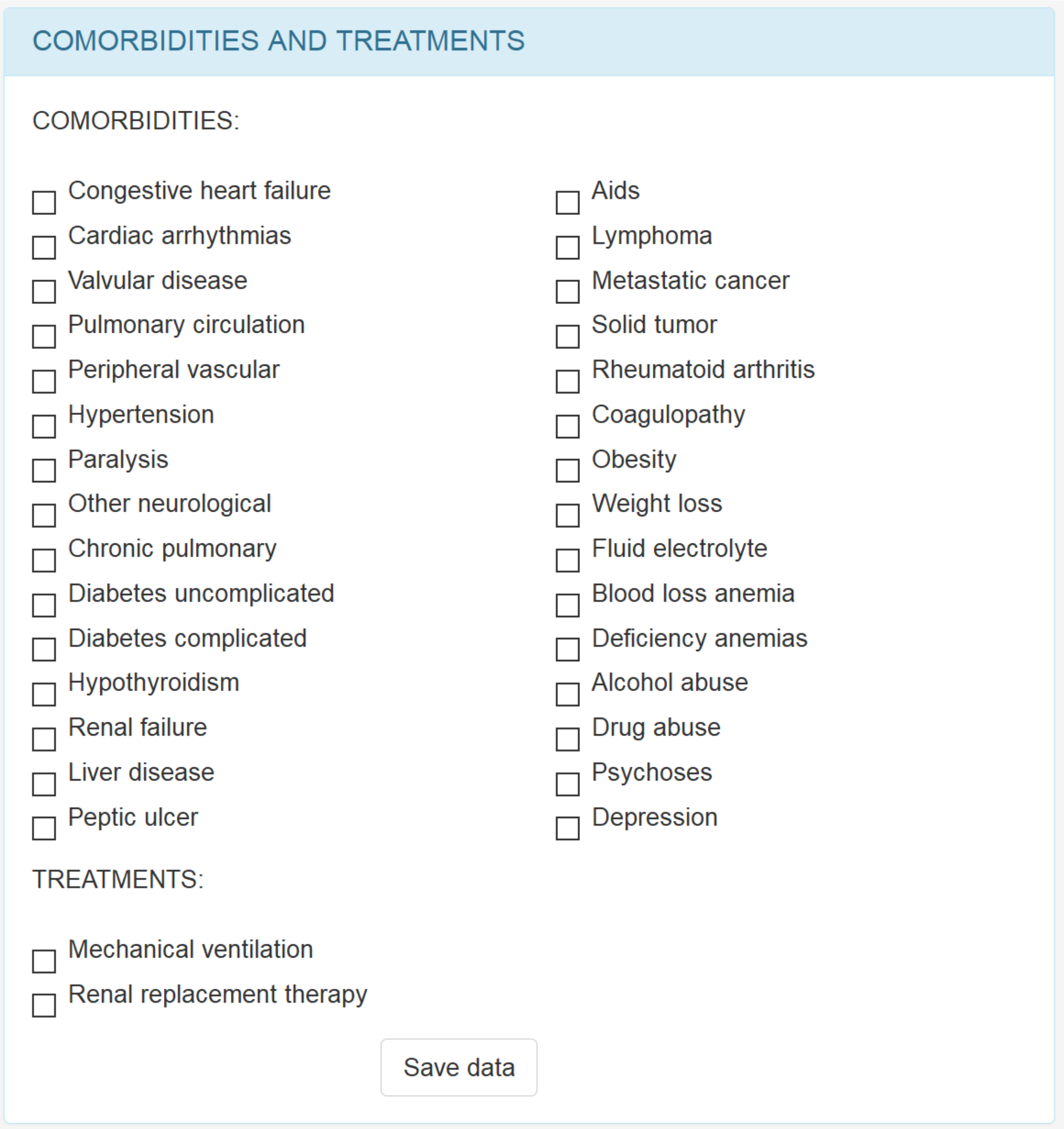
Birthdate: 2019-05-08

Gender: Male

Admission type: Emergency

Figure 11.4. Admission data and demographics collapsible panel.

The comorbidities and treatments collapsible panel is composed of 30 checkboxes based on the Elixhauser Comorbidity groups and two checkboxes for two treatments dialysis and mechanical ventilation). Figure 11.5 presents the final appearance of the comorbidities and treatments collapsible panel.



COMORBIDITIES AND TREATMENTS

COMORBIDITIES:

□ Congestive heart failure
□ Cardiac arrhythmias
□ Valvular disease
□ Pulmonary circulation
□ Peripheral vascular
□ Hypertension
□ Paralysis
□ Other neurological
□ Chronic pulmonary
□ Diabetes uncomplicated
□ Diabetes complicated
□ Hypothyroidism
□ Renal failure
□ Liver disease
□ Peptic ulcer

□ Aids
□ Lymphoma
□ Metastatic cancer
□ Solid tumor
□ Rheumatoid arthritis
□ Coagulopathy
□ Obesity
□ Weight loss
□ Fluid electrolyte
□ Blood loss anemia
□ Deficiency anemias
□ Alcohol abuse
□ Drug abuse
□ Psychoses
□ Depression

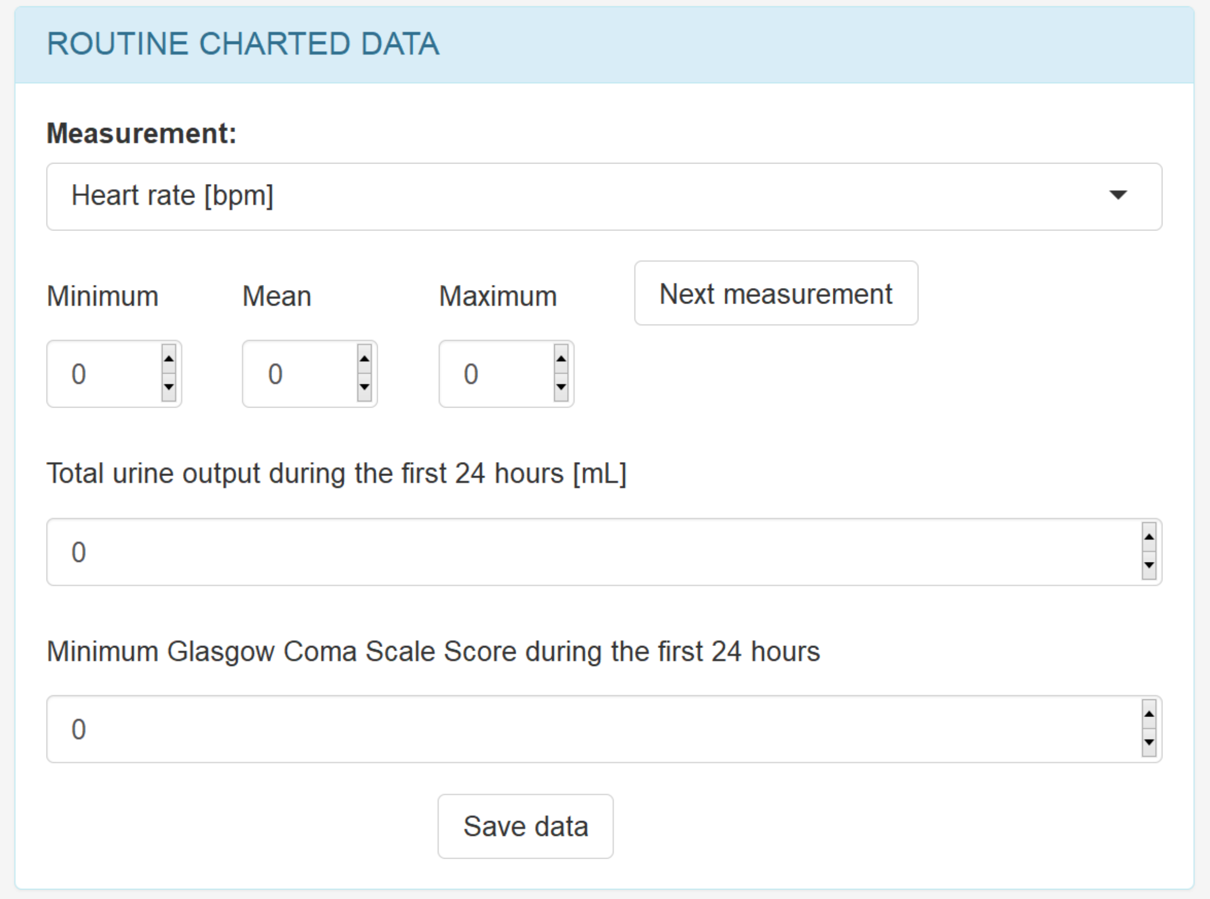
TREATMENTS:

□ Mechanical ventilation
□ Renal replacement therapy

Figure 11.5. Comorbidities and treatments collapsible panel.

The routine charted data collapsible panel contain two elements, in the first one there is a selection input for eight routine charted measures (Heart rate, Arterial Blood Pressure Systolic, Arterial Blood Pressure Diastolic, Arterial Blood Pressure Mean, Respiratory rate, Temperature, Peripheral capillary oxygen saturation and Glucose) and three numeric inputs for the minimum, mean and maximum values of each measurement, each time the user clicks on the "Next measurement" button, the system automatically stores the values of the current measurement in the feature vector and changes the selection to the next measurement.

The second element of the routine charted data collapsible panel is composed of two numeric inputs for the total urine output during the first 24 hours and the minimum Glasgow Coma Scale Score during the first 24 hours. Figure 11.6 present the final appearance of the admission data and demographics collapsible panel.



ROUTINE CHARTED DATA

Measurement:

Heart rate [bpm]

Minimum: 0

Mean: 0

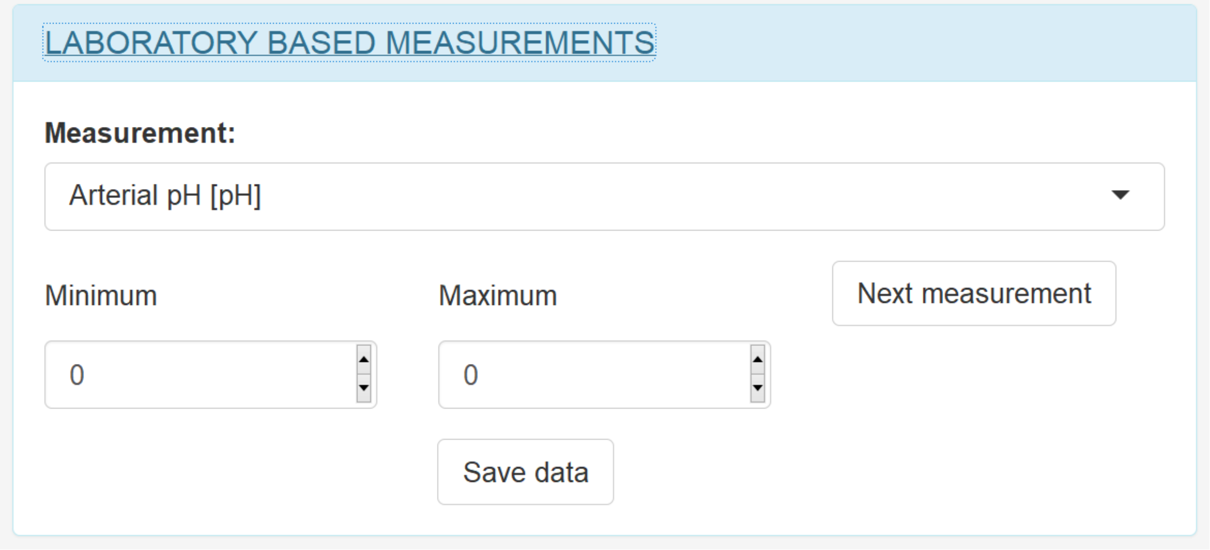
Maximum: 0

Total urine output during the first 24 hours [mL]: 0

Minimum Glasgow Coma Scale Score during the first 24 hours: 0

Figure 11.6. Routine charted data collapsible panel.

The laboratory based measurements collapsible panel contain a selection input for 17 laboratory test (Arterial pH , Anion gap , , Bilirubin , Creatinine , Chloride , Hematocrit, Hemoglobin , Lactate , Platelet Count , Potassium , Partial thromboplastin time (PTT, International normalized ratio (INR), Prothrombin time (PT) , Sodium , Blood urea nitrogen (BUN) , White Blood Cell (WBC) count) and two numeric inputs for the minimum and maximum values of each measurement, each time the user clicks on the "Next measurement" button, the system automatically stores the values of the current measurement in the feature vector and changes the selection to the next measurement. Figure 11.7 present the final appearance of the laboratory based measurements collapsible panel.



LABORATORY BASED MEASUREMENTS

Measurement: Arterial pH [pH]

Minimum: 0

Maximum: 0

Figure 11.7. Laboratory based measurements collapsible panel.

The visualization panel contains three elements, the most important one is the one-year mortality probability, which is obtained from a personalized SGB model and presented in a text. The second one is the mortality rate among the 100 most similar patients, this result is presented in a pie chart. The final one is the relative importance of the ten most relevant variables for the precision cohort selected from the new patient, this information is presented in a column chart and it is obtained with the SGB variable ranking. Figure 11.8 present the final appearance of the output panel.

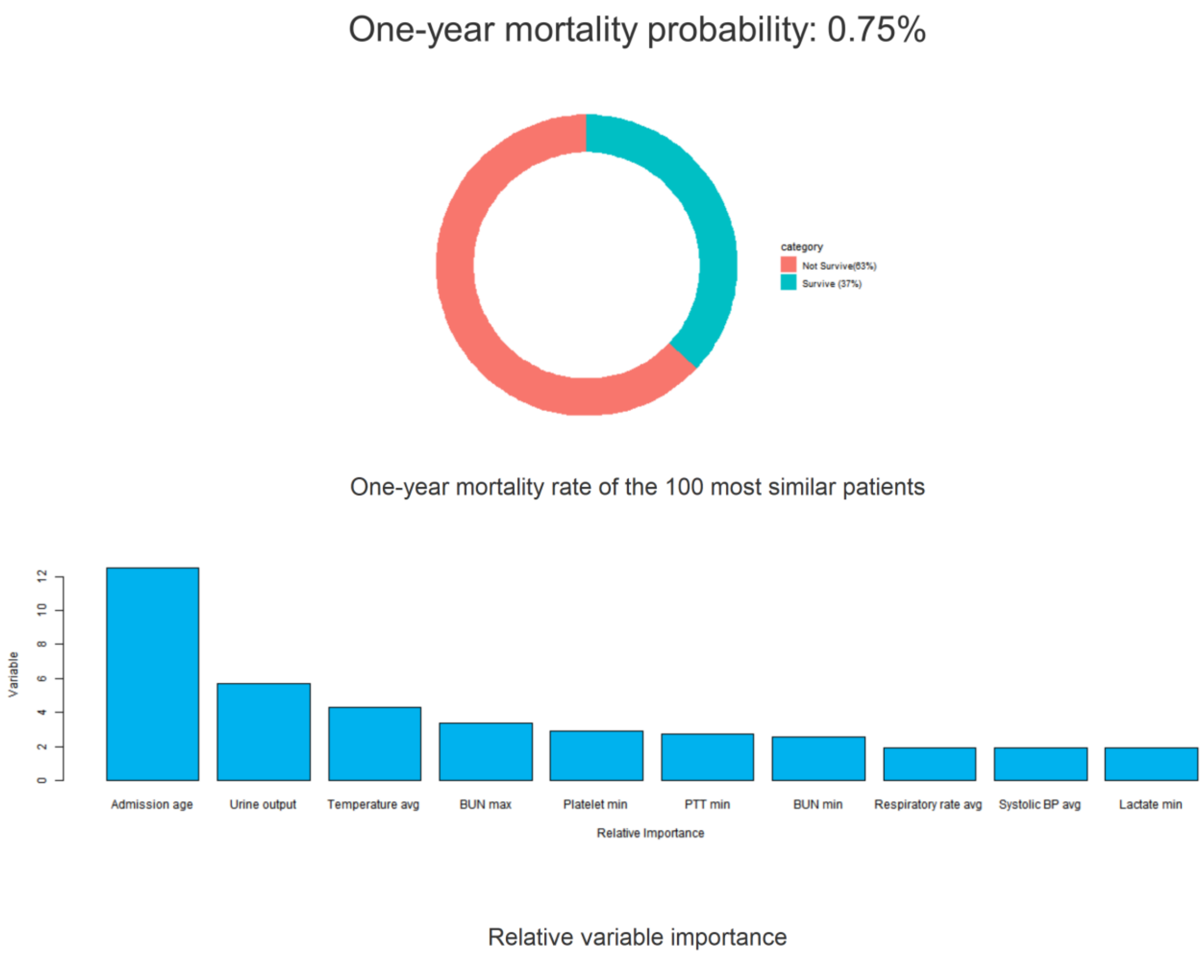
11.4 Conclusions

The creation of software enables the clinical use of machine learning models developed for the prediction of one-year mortality of sepsis patients within the intensive care unit.

The three outputs are personalized for each new patient, since they are based on the Weighted Contribution Similarity, a measure that proved to generate model with better discrimination capability that the model fitted with the entire study cohort.

The one-year mortality rate among the 100 most similar patients and the top ten most relevant variables, help to provide context about the patient's condition, which can be used by medical personnel to complement their diagnosis and provide better treatment.

The use of a tool like Shiny for the development of a web application ensures that the software can be used from any operating system, in addition, the fact that this tool has integrated functions for scaling the content according to the size of the screen allows its use in mobile devices.


One-year mortality probability: 0.75%

One-year mortality rate of the 100 most similar patients

Relative variable importance

Figure 11.8. Output panel.

PART 5: SUMMARY, CONCLUDING REMARKS, DIFFUSION AND FUTURE PERSPECTIVES

This part present the final considerations of our study, we present a summary of the process carried out in this thesis, we indicate what were our main achievements and we show the limitations of our work and the future perspectives derived from them.

CHAPTER 12. SUMMARY, CONCLUDING REMARKS, DIFFUSION AND FUTURE PERSPECTIVES

12.1 Summary

General severity of illness scores can be useful to several purposes: guide prognostication, to assess ongoing disease development and organ function, to compare ICU performance over time and across units, to compare clinical trial population and outcomes. In a survey Bouch listed the characteristics for an ideal scoring system [33]:

1 On the basis of easily/routinely recordable variables
2 Well calibrated
3 A high level of discrimination
4 Applicable to all patient populations
5 Can be used in different countries
6 The ability to predict functional status or quality of life after ICU discharge.

None of the current scoring system incorporates all these features; Moreover, items 4 and 5 are challenging to fulfill, that is why customize scoring systems, are increasingly being developed. The customize models have proved to perform better than the general population approaches, however these studies continue to be population-based and therefore they generally provide “the average best choice”.

Personalized mortality predictive modeling based on patient similarity is a developing field that seeks to identify patients who are similar to an index (new) patient and derive insights from the data of similar patients to provide personalized predictions. This approach has been widely used for personalized predictions in other fields, including music, movies and e-commerce, however, there are still very few studies that focus on personalized prediction driven by patient similarity metrics within the ICU.

In the specific case of sepsis, a condition associated with ongoing mortality beyond short-term end points (i.e. in-hospital mortality), and additional morbidities such as higher risk of readmissions, cardiovascular disease and cognitive impairment for survivors; specific models for the mortality prediction within the ICU have been developed, which presented better performance than adjusted traditionally severity of disease scores, however, these sepsis-customized models focus on the short term mortality prediction (7-day mortality and in-hospital mortality). Studies suggest that the use of in-hospital mortality as an end point for clinical studies are not enough to understand the effect of sepsis on mortality and quality of life, and the current understanding of the risk factors and mechanisms underlying long-term sequelae in patients that suffered from this condition still limited. Therefore, identify risk factors during an ICU stay that reverberate and even could predict long-term outcomes will help physicians offer better treatments.

The first step to build a long-term mortality prediction model for sepsis patients within the ICU is to obtain quality data. It is clear, that the performance of the models depends on the characteristics of the used database because machine learning techniques will give poor performance, lead to imprecise and inaccurate conclusions or even fail to find a good predictive model if the database is too noisy or if it is not representative of the studied population. Therefore, high-quality clinical databases are of value in clinical practice, in managing services and in developing health technologies.

The selected database for this work is MIMIC-III which data comes from a single institution (Beth Israel Deaconess Medical Center in Boston Massachusetts). However, despite the limitation of being single-centered, the main advantages of MIMIC-III are:

- Right now the only freely accessible critical care database of its kind.
- The dataset spans more than a decade.
- It has detailed information about individual patient care that includes time-stamped nurse-verified physiological measurements and out-of-hospital mortality dates.

From the MIMIC-III clinical database we extracted two study cohorts, the first one composed of patients who presented an explicit sepsis diagnosis or fulfill the Angus criteria; and the other conformed with patients that meet, besides the explicit sepsis and Angus criteria, the Martin and Sepsis-3 criteria. In order to develop the models, we included basic descriptors (like the minimum, maximum and mean) for each of the numerical continuous variable (vital signs, laboratory measurements and admission age), however, we do not include measures of the tailedness of the distributions (like kurtosis or skewness) for two reasons, first, their inclusion would affect the interpretability of the models, and second, the clinical operatization of ICU stay in Colombia do not ensure storage of the data at an adequate temporal resolution to obtain reliable values of these descriptors, that is to say, the medical devices do not usually automatically register the values of the vital signs as it is done in the Beth Israel Deaconess Medical Center in Boston Massachusetts.

According to the above, we focus on the development of a model that can be used to patient individual prognostication and goes beyond the prediction of in-hospital mortality. For this, we divide this project in four different stages that, progressively, lead to the generation of models that can be used for the individual one-year mortality prediction of patients with sepsis within the intensive care unit clinical practice.

In the first stage, presented in **chapter 6,** we developed a customized Stochastic Gradient Boosting (SGB) model for the one-year mortality prediction and compared it performance with three adjusted models based on traditional severity of disease scoring systems. In this stage we used a study cohort composed of 5650 sepsis patients which presented an explicit sepsis diagnosis or fulfill the Angus criterion for sepsis. Our customized prediction model proved to outperform adjusted traditional severity of disease scores since it obtained an AUROC of 0.805 (95% Confidence Interval: 0.785 - 0.826) and the best performing reference model was the adjusted SAPS II model that obtained a AUROC of 0.702 (95% Confidence Interval: 0.683 - 0.719). Besides this, in this stage we also obtained a subset of predictors truly related with one-year mortality which were selected using the Least Absolute Shrinkage and Selection Operator and the SGB variable importance ranking. This subset of predictors was composed of 17 variables which are the admission age the following values from the first 24 hours of admission: total urine output, blood urea nitrogen maximum, lactate minimum, hemoglobin mean, temperature maximum, glucose minimum, temperature minimum, spo2 mean, bilirubin maximum, platelet count maximum, systolic blood pressure maximum, white blood cell count minimum and the following comorbidities and treatment: metastatic cancer, malignancy, hypertension and mechanical ventilation. The subset of the 17 variables allowed the creation of a simpler SGB model that maintained the good performance since it reported an AUROC of 0.791 (95% Confidence Interval: 0.769 - 0.812). The calibration of the developed customized SGB models and the adjusted traditional severity of disease scoring systems were also evaluated; and we found that all the developed SGB models presented an adequate calibration but of the adjusted reference models, only SOFA presented an adequate calibration.

In the second stage we developed two customized scores for the stratification of patients in risk groups, which is presented in **chapter 7**. In this stage we focus on the subset of 17 variables that were found to be relevant in the previous stage and complement them with variables that are frequently used within the ICUs. In addition to this, we increase the criteria for identifying patients with sepsis, and we include those that meet the Martin and sepsis-3 criteria. As a result, we gathered a study cohort composed of 15082 admissions. With this study cohort, we generated two scoring systems for the assessment of the one-year mortality risk of sepsis patients within the ICU. The first score was based on dichotomization of the variables and achieved an AUROC of 0.769 (95% Confidence Interval: 0.761 - 0.778) on a validation subset composed of 9049 admissions; the second generated score used multiple cutoff points for each continuous numerical variable (i.e. the laboratory measurements, the routine charted data and the admission age) and achieved an AUROC of 0.785 (95% Confidence Interval: 0.783 - 0.794). Although the multiple Cutoff points score presented better discrimination and it use will allow a more accurate interpretation of the condition of each patient, the binary CP is more easy to implement and can be used for a quick interpretation of a patient's condition. These developed scores presented adequate calibration and outperformed adjusted traditional severity of disease classification systems over the same validation subset.

The third stage was the personalized predictive modeling based on patient similarity. We used a lineal approach, presented in **chapter 8**; and a non-lineal approach presented in **chapter 9**. The development of the personalized one-year mortality prediction model is based on patient similarity measures and follows the next outline:

5) All pairwise Similarity measures between the index admission and every admission in the training data were calculated. In **chapter 8** five different measures that model the interaction between admissions were developed and evaluated, in **chapter 9** we only used the weighted contribution similarity, a patient similarity metric based on the fact that different conditions carry different mortality risk; the weights for each condition was obtained from the scores developed in the previous stage.
6) The calculated similarity values were sorted in ascending order.
7) A precision cohort was created with the data of the n most similar admissions. In chapter 8 the number of most similar admission was varied from 1000, to 13000; in chapter 9, the number of most similar admissions were settled to 4000, since with this number the peak mean AUROC were obtained.
8) Each precision cohort was used to train a personalized mortality prediction model for the index admission. In chapter 8 we used logistic regression model, and in chapter 9 we used SGB. The mean AUROC of the personalized SGB models were 0.809 (95% Confidence Interval: 0.791 - 0.825)

and the mean AUROC of the personalized Logistic regression models were 0.794 (95% Confidence Interval: 0.78 - 0.807).

In chapter 10, we evaluated an approach framed as graph-based learning, where label information is smoothed over the graph via some form of explicit graph-based regularization. For this we construct a graph based on the weighted contribution patient similarity metric with the same weights that the previous chapter and use it to train a regularized multilayer neural network model, which reported a peak AUROC of 0.812 (95% Confidence Interval: 0.808 - 0.815).

In the final stage, presented in **chapter 11**, we developed a software that could be used in the clinical environment, this software is based on the characteristics of stages 2 and 3, that most contributed to the discrimination of the long-term mortality of patients with sepsis within the ICU, and provide information of clinical utility, according to the criterion of intensivist experts.

12.2 Concluding Remarks

The approach presented in **chapter 6**, showed that SGB variable importance and LASSO methodologies allowed the identification of a subset of predictors that are significatively related to the one-year mortality prediction of sepsis patients within the ICU. The SGB models developed with only the variables selected with either of those methods preserved the same performance as the one generated with all the predictors. Also the intersection of the predictors selected by the two methods lead to the development of a much simpler model with only 17 predictors, that also presented a similar performance to the complete model.

The main objective of this stage was to present a customized model for the one-year mortality prediction of the patients that are admitted in a ICU with a sepsis diagnosis; and shows that the use of ensemble based algorithms (SGB) and the inclusion of predictors that are not usually taken into account in the traditional severity-of-disease classification systems (for example minimum lactate), improves the performance of the prediction of prognosis models in patients admitted to an ICU with diagnosis of sepsis.

In the second stage, presented in **chapter 7**, we present the assessment of two customized severity-of-illness scoring system specifically for patients with sepsis. The scores utilized 6033 admissions for its development and 9049 for its validation. The first score was based on the dichotomization of the numerical continuous variables; the second score is based on multiple cutoff points for each numerical continuous variable. Both scores accurately estimated the probability of one-year mortality in sepsis diagnosed patients within the ICU and were well calibrated.

The strengths of this scores are that they performed well with respect to both discrimination and calibration. The calibration is especially important as data were collected from a database that spams over 10 years and we used four different sepsis criteria to retrospectively identify the admissions, including sepsis-3 that is the most recent one. The scores are composed of 52 (for the binary cutoff points score) and 92 variables (for the multiple cutoff points score). The number of variables are more than traditional scores like SAPSII (20 variables) and SSS (36 variables) but considerably less than most recent approaches like APACHE IV (142 variables). The objective of this score is to early alert of a worse prognostic and to stratify patients according to their risk.

AUROC analysis showed that the developed scores outperformed adjusted traditional severity of disease scoring systems, even more, the predictive capacity of the score is better, in this study, than the SOFA, that is the scoring systems used for the most recent sepsis and septic shock consensus [17, 18], and the Sepsis Severity Score (SSS), that is an internationally derived scoring system specifically for patients with severe sepsis and septic shock, however, this scores continue to be a population-based approach and therefore they provide “the average best choice” for sepsis patients. For this reason, we focused on the developing of personalized predictive models based on patient similarity.

In **chapter 8** the utility of similarity metrics in personalizing one-year mortality risk estimation in the ICU for sepsis patients was proved. The results showed that using a subset of similar admission rather than a larger population as training data improves one-year mortality prediction performance, even when the population shares a common characteristic.

Although all the evaluated admissions are from patients with sepsis (which means that they all have an infection and an organ dysfunction), there was improvement when using similarity metrics, even more a simple mortality rate among 100 similar admissions resulted in good predictive performance that exceeded the performance obtained with the adjusted traditional severity of disease scoring systems.

Traditionally the risk prediction in a ICU, is addressed by the clinician based on large population studies of patients, like severity of disease scores, however the developed models outperformed widely the adjusted traditional used scores, which could be explained by the following elements:

- The use of a specific cohort of patients with sepsis.
- The inclusion of sepsis related variables, like lactate.
- The fact that nearby admissions are more comparable and tend to have the same outcome.

The first two elements explain why the logistic regression model fitted with all the training subset exceeded the performance of the currently used scores, the third one explains the improvements observed when personalized models based on the precision cohort of each patient are used.

An important aspect of the personalized logistic regression approach is that it gives particular coefficients for each precision cohort which could be interpreted as relative variable importance for a particular patient, meaning that a treating doctor could elucidate the most relevant factor in the prediction, so it has the potential to provide tailored prognoses, and prescribe more effective treatments.

It is clear that one of the factors that strongly affects predictive performance is the choice the similarity measure, for this reason, five similarity measures were tested, of which the weighted contribution similarity with Scoring System for the One-Year Mortality Prediction of Sepsis Patients weights generated the personalized one-year mortality prediction models with better predictive performance; however the other weighted contribution similarities did not show a better performance than Equal contribution similarity, one possible explanation could be that the scores from which these weights were derived were developed for in-hospital mortality and they do not include all the considered comorbidities and treatments.

Despite the good results for in-hospital mortality prediction using a cosine similarity based approach reported by Lee et Al. [2], in this study, cosine similarity was, in general, the one that had a worse performance, indicating that importance of the comorbidities when evaluating the long-term mortality. Moreover, the predictive performance improvement reported by Lee et Al. was 2.47% (the best performing model presented an AUROC of 0.83 and the model that used all available data for training presented an AUROC of 0.81), but predictive performance improvement in our study was 1.15%. This could be explained by the fact that the population that we evaluated is especially homogeneous; all of the patients in our study cohort present the same severe diagnosis, sepsis, have a median ICU length of stay of 4 days, a median hospital length of stay of 11 days and a median age of 68 years old.

In **chapter 8** we demonstrated the value of patient similarity-based models in critical health problems and shows the superiority of patient similarity-based models over population-based ones. In order to improve the discrimination between those patients who survive more than one year after ICU sepsis-related admission and those who do not, in **chapter 9** we implemented personalized Stochastic Gradient Boosting models and in **chapter 10** we developed a graph-based Laplacian regularized multilayer neural network. The main difference between those approaches is that the first one creates a particular model for each new patient; and the second generates a single model that is generalizable to unobserved instances, which mean that it is only trained one time.

Best overall performance was obtained with the Laplacian regulated neural network, however since it is based on a three-layer neural network loses interpretability. An important factor in favor of the personalized logistic regression approach is that it could give a relative variable importance for each precision cohort, so it has the potential to provide tailored prognoses, and prescribe more effective treatments.

In the section 1.3 in **chapter 1**, we reported some of the most relevant studies in the field of prediction of mortality within the ICU and the main conclusion were:

- Ensemble methodologies based on trees consistently report good performances.
- Mortality prediction can be approached quite linearly.
- Deep learning models require large training and feature sets to report improvements,
- Selecting an appropriate similarity metric is not a straightforward task.

The first two items were addressed and proved in this thesis since, in chapter 6 we developed a model based on stochastic gradient boosting (SGB), an ensemble tree methodology, that outperformed the reference adjusted models based on severity scores, and in chapter 7, we presented a scoring system, based on logistic regression, that accurately indicates the risk of one-year mortality prediction of sepsis patients admitted to the ICU.

The third item was glimpsed in the fact that no significant difference is observed between the Laplacian regulated neural network and the personalized SGB models, the foregoing suggests that a future path of investigation through the ways of deep learning will require a larger study population, and a greater number of input variables.

The fourth item, it is in which we have our greatest contribution, since we managed to generate a similarity metric that prove to be relevant to personalized prediction models based on logical regression, but that also helps to improve the performance of non-linear approaches such SGB and the Laplacian regulated neural network.

In synthesis, the discrimination analysis over the models presented in chapter 6 and chapter 7 indicate that customized mortality prediction models for a specific disease presents a better performance that traditional scores; and the personalized models developed in chapter 8, chapter 9 and chapter 10 surpass the performance of population-based models; moreover, the results presented shows that this thesis is methodologically comparable to the state-of-the-art machine learning approaches to the outcome prediction problem, and specifically in the field of personalized mortality prediction models represents an advance in the state of the art, since we achieve a similarity metric that improves the performance of both linear and nonlinear models.

Despite the good performance of the models developed, it is clear that, each of these models are difficult to interpret; therefore, it is necessary to develop easy-to-use computer tools that allow these types of models to be implemented within the ICU.

In **chapter 11** we presented the creation of a software that enables the clinical use of the machine learning models developed for the prediction of one-year mortality of sepsis patients within the intensive care unit. The software present three outputs that are personalized for each new patient, since they are based on the Weighted Contribution Similarity, a measure that proved to generate model with better discrimination capability that the model fitted with the entire study cohort.

From this outputs, the one-year mortality rate among the 100 most similar patients and the top ten most relevant variables, help to provide context about the patient's condition, which can be used by medical personnel to complement their diagnosis and provide better treatment; and the one-year mortality probability indicates the individual risk of each patient according to their precision cohort.

Wrapping up, in this thesis we developed models that successfully identify those patients who are at risk of dying one year after their sepsis related admission using demographic variables, comorbidities and physiological data obtained during the first 24 hours of their ICU stay. The clinical usefulness of this is immense since it has been proved that patients with sepsis have ongoing mortality beyond short-term end points, and survivors consistently demonstrate impaired quality of life, and models that the ones presented in this thesis allow the early identification of those patients at higher one-year mortality risk, therefore, these patients could be observed attentively and they could be given additional care that will improve their quality of life.

The customized scores presented in **chapter 7** generate a segmentation of the sepsis patients in five groups according to their one-year mortality risk. Since the customized scores were developed using exclusively the data from patients with ICU sepsis-related admissions, they have better discriminatory ability than the adjusted models based on traditional scores, moreover, since our output variable focuses on long-term mortality, the presented customized models also outperform models that are been currently used on the sepsis population, such as Sequential [Sepsis-related] Organ Failure Assessment (SOFA) score, and Severe Sepsis Mortality Prediction Model (SSS).

The customized models presented in **chapter 6** and **chapter 7**, allow the assessment of the one-year mortality probability of sepsis patients within the ICU, however, methodologies were also developed that allowed the generation of personalized models which can be used at the patient level.

12.3 Limitations and future Perspectives

This thesis shows that it is possible to identify those who are at risk of dying one year after their sepsis related admission using demographic variables, comorbidities and physiological data obtained during the first 24 hours of their ICU stay, moreover we have proven that patient similarity metrics can improve discrimination ability. However, our work has certain limitations; First of all, the developed models are based on the admissions taken from a single single-center, however, we consider that the final sample is sufficiently representative to generalize the results, since the admissions included in the study cohort were obtained using multiple retrospective sepsis identification criteria and the used database spans for more than 10 years.

Another important limitation, is that the personalized models, requires a relatively high among of time to identify the precision cohort for a new patient, concretely, the methodology used in **chapters 8 and 9** implies the construction of a similarity matrix $W \in \mathbb{R}NxN$ where N is the total number of observations, which implies an algorithm with complexity $O(n^2)$, moreover, for the identification of the precision cohort, an sorting function is also used over the vector with the similarity metric values between any patient and the remaining N-1 patients, a process that implies an algorithm with complexity $O(n \log n)$.

According to the above, it is clear that, with current implementations, the number of instructions required to identify the precision cohort would increase drastically if the size of the population increases, therefore, the research and implementation of algorithms that have a linear complexity $O(n)$, and the adaptation of the developed methodologies to parallel and distributed computing environments is proposed as future work. Nevertheless, we consider that, with our current population, such time is not excessive (about 20 minutes in a regular laptop, but not formal test was executed) considering the fact that the model can be applied just 24 hours after the ICU admission and it serves to estimate the probability of mortality one year after the admission.

It is important to note that for the non-linear models presented in this thesis (personalized SGB and graph-based regularized multilayer neural network) a much smaller amount of tests runs were executed compared to linear models; The reason for this was the large amount of computing time necessary to evaluate these models on more than a thousand patients. Similarly, the parameters used for the identification of the precision cohort in the non-linear models were those that presented the best performance when the personalized logistic regression models were constructed, and it is reasonable to expect that different algorithms have different optimal parameters, in particular, we consider that the adequate number of similar patients can be especially sensitive to the used machine learning technique; Unfortunately, with the computing capacity and the algorithms that we currently have, these tests were considered unfeasible.

According to the above it would be desirable to externally evaluate the generated models in a Colombian context, for this it is necessary to build a quality database which would imply technical challenges associated with the data acquisition and storage, procedural challenges associated with clinical permits and ethical challenges related to patient privacy.

Our results indicate that a well selected patient similarity metric improves discrimination ability, for this reason further developments in the field of personalized models within the ICU should focus on the selection of a good similarity measure; according to this, in the short term techniques like feature selection, predictor weighting schemes, or experts' opinions should be used to develop new patient similarity measure that improve the performance of prediction models. It will be also interesting to implement distance metric learning approaches.

On the other hand, our graph-based regularized multilayer neural network prove to be a promising route, thus, novel machine learning approaches based on graph-structured data like graph convolutional networks should be implemented for mortality prediction task.

Finally, in order to reduce the computation time, it would be necessary to implement the methodologies presented in this thesis on a software framework that allow distributed storage and processing of data such as Apache Spark.

12.4 Diffusion: Publications and Conference presentation

Parts of the work presented in this thesis have been published in international journals and the proceedings of international conferences. Publications relating to this work are listed below:

1. García-Gallo, J. E., Fonseca-Ruiz, N. J., Celi, L. A., & Duitama-Muñoz, J. F. (2018). A machine learning-based model for 1-year mortality prediction in patients admitted to an Intensive Care Unit with a diagnosis of sepsis. Medicina intensiva.
2. García-Gallo, J. E., Fonseca-Ruiz, N. J., & Duitama-Muñoz, J. F. (2019). Scoring System for the One-Year Mortality Prediction of Sepsis Patients in Intensive Care Units. In World Congress on Medical Physics and Biomedical Engineering 2018 (pp. 367-370). Springer, Singapore.

Presentations were also made at national and international conferences:

1. III Seminario Internacional de Actualización en Ingeniería Biomédica. Held in Bucaramanga, Colombia from octuber 26 to 27, 2017.
2. IEEE Conference on Biomedical and Health Informatics (BHI) 2018. held in Las Vegas, United States of America, from March 4 to 7, 2018.
3. 10th International Conference on Bioinformatics and Biomedical Technology (ICBBT 2018) held in Amsterdam, Netherlands, from May 16 to 18, 2018.
4. World Congress on Medical Physics and Biomedical Engineering 2018 (WC2018) held in Prague, Czech Republic, from June 3 to 8, 2018.

REFERENCES

1. Celi LA, Galvin S, Davidzon G, et al (2012) A Database-driven Decision Support System: Customized Mortality Prediction. J Pers Med 2:138–48 . doi: 10.3390/jpm2040138

2. Lee J, Maslove DM, Dubin J a (2015) Personalized mortality prediction driven by electronic medical data and a patient similarity metric. PLoS One 10:e0127428 . doi: 10.1371/journal.pone.0127428

3. Takrouri MSM (2004) Intensive care unit. Internet J Heal 3:2–4

4. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) APACHE II: a severity of disease classification system. Crit Care Med 13:818–829

5. Le Gall J-R, Lemeshow S, Saulnier F (1993) Simplified Acute Physiology Score (SAPS II) Based on a European / North American multicenter study. Jama 270:2957–2963 . doi: 10.1001/jama.270.24.2957

6. Johnson AEW (2014) Mortality prediction and acuity assessment in critical care

7. Vincent JL, Moreno R (2010) Clinical review: scoring systems in the critically ill. Crit Care 14:207 . doi: 10.1186/cc8204

8. Rapsang A, Shyam D (2014) Scoring systems in the intensive care unit: A compendium. Indian J Crit Care Med 18:220 . doi: 10.4103/0972-5229.130573

9. Knaus WA, Zimmerman JE, Wagner DP, et al (1981) APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 9:591–597

10. Zimmerman JE, Kramer AA, McNair DS, Malila FM (2006) Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. Crit Care Med 34:1297–1310 . doi: 10.1097/01.CCM.0000215112.84523.F0

11. Moss TJ, Calland JF, Enfield KB, et al (2017) New-Onset Atrial Fibrillation in the Critically Ill*. Crit Care Med 45:790–797 . doi: 10.1097/CCM.0000000000002325

12. Johnson AEW, Kramer AA, Clifford GD (2013) A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy. Crit Care Med 41:1711–1718 . doi: 10.1097/CCM.0b013e31828a24fe

13. Johnson AEW (2014) Mortality prediction and acuity assessment in critical care. University of Oxford

14. Zhang M, Wang Z, Hong X, et al (2017) The study of the value of Oxford Acute Severity of Illness Score in assessing the severity of critical illness patients: a single-center analysis of 470 cases. Chinese J Emerg Med 26:197–201

15. Le Gall J-R, Klar J, Lemeshow S (1997) How to assess organ dysfunction in the intensive care unit? The logistic organ dysfunction (LOD) system. Sepsis 1:45–47

16. Vincent J-L, Moreno R, Takala J, et al (1996) The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Med 22:707–710

17. Opal SM, Rubenfeld GD, Poll T Van Der, et al (2016) The Third International Consensus Definitions

for Sepsis and Septic Shock (Sepsis-3). JAMA J Am Med Assoc 315:801–810 . doi: 10.1001/jama.2016.0287

18. Seymour CW, Liu VX, Iwashyna TJ, et al (2016) Assessment of Clinical Criteria for Sepsis. Jama 315:762 . doi: 10.1001/jama.2016.0288

19. Hu Y, Zhang X, Liu Y, et al (2013) APACHE IV Is Superior to MELD Scoring System in Predicting Prognosis in Patients after Orthotopic Liver Transplantation. Clin Dev Immunol 2013:5

20. Haq A, Patil S, Parcells AL, Chamberlain RS (2014) The Simplified Acute Physiology Score III Is Superior to the Simplified Acute Physiology Score II and Acute Physiology and Chronic Health Evaluation II in Predicting Surgical and ICU Mortality in the "' Oldest Old .'" 2014:10–13

21. Saleh A, Ahmed M, Sultan I, Abdel-lateif A (2015) Comparison of the mortality prediction of different ICU scoring systems (APACHE II and III, SAPS II, and SOFA) in a single-center ICU subpopulation with acute respiratory distress syndrome. Egypt J Chest Dis Tuberc 64:843–848 . doi: 10.1016/j.ejcdt.2015.05.012

22. Arabi Y, Shirawi N Al, Memish Z, et al (2003) Assessment of six mortality prediction models in patients admitted with severe sepsis and septic shock to the intensive care unit: a prospective cohort study. Crit care 7:R116–R122 . doi: 10.1186/cc2373

23. Sun D, Ding H, Zhao C, et al (2017) Value of SOFA, APACHE IV and SAPS II scoring systems in predicting short-term mortality in patients with acute myocarditis. Oncotarget 8:63073–63083 . doi: 10.18632/oncotarget.18634

24. Jentzer JC, Murphree DH, Wiley B, et al (2018) Comparison of Mortality Risk Prediction Among Patients ≥70 Versus <70 Years of Age in a Cardiac Intensive Care Unit. Am J Cardiol 1–6 . doi: 10.1016/j.amjcard.2018.08.011

25. Johnson AEW, Dunkley N, Mayaud L, et al (2012) Patient Specific Predictions in the Intensive Care Unit Using a Bayesian Ensemble. Comput Cardiol (2010) 39:249–252

26. Lee J, Maslove DM (2015) Customization of a Severity of Illness Score Using Local Electronic Medical Record Data. J Intensive Care Med

27. Pirracchio R, Petersen ML, Carone M, et al (2015) Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): A population-based study. Lancet Respir Med 3:42–52 . doi: 10.1016/S2213-2600(14)70239-5

28. Johnson AEW, Pollard TJ, Mark RG (2017) Reproducibility in critical care: a mortality prediction case study. Proc Mach Learn Healthc 68:361–376

29. Lee J (2017) Patient-specific predictive modeling using random forests: an observational study for the critically ill. JMIR Med informatics 5:e3

30. Purushotham S, Meng C, Che Z, Liu Y (2018) Benchmarking deep learning models on large healthcare datasets. J Biomed Inform 83:112–134 . doi: 10.1016/j.jbi.2018.04.007

31. Barrett LA, Payrovnaziri SN, Bian J, He Z (2018) Building Computational Models to Predict One-Year Mortality in ICU Patients with Acute Myocardial Infarction and Post Myocardial Infarction Syndrome. 1:407–416

32. Bice T, Carson SS (2017) Prolonged mechanical ventilation. Evidence-Based Crit Care A Case Study

Approach 40:251–256 . doi: 10.1007/978-3-319-43341-7_28

33. Bouch CD, Thompson JP (2008) Severity scoring systems in the critically ill. Contin Educ Anaesthesia, Crit Care Pain 8:181–185 . doi: 10.1093/bjaceaccp/mkn033

34. Gül F, Arslantaş MK, Cinel İ, Kumar A (2017) Changing definitions of sepsis. Turk Anesteziyoloji ve Reanimasyon Dern Derg 45:129–138 . doi: 10.5152/TJAR.2017.93753

35. Chen H, Meigs JB, Division M, et al (2015) HHS Public Access. 78:81–90 . doi: 10.1016/S0140-6736(12)61815-7.Sepsis

36. Bennett SR (2015) Sepsis in the intensive care unit. Surg 33:565–571

37. Sibbald WJ, Doig G, Inman KJ (1995) Sepsis, SIRS and infection. Intensive Care Med 21:299–301

38. Banerjee D, Levy MM (2017) Sepsis Definitions. In: Sepsis. Springer, pp 7–24

39. Shankar-Hari M, Phillips GS, Levy ML, et al (2016) Developing a New Definition and Assessing New Clinical Criteria for Septic Shock. Jama 315:775 . doi: 10.1001/jama.2016.0289

40. Singer M, Deutschman CS, Seymour CW, et al (2016) The third international consensus definitions for sepsis and septic shock (sepsis-3). Jama 315:801–810

41. Singer M (2017) The new definitions of SEPSIS and SEPTIC SHOCK: What do they give us? An answer. Med Intensiva 41:41–43 . doi: 10.1016/j.medin.2016.10.015

42. Rodríguez A, Martín-Loeches I, Yébenes JC (2016) New definition of sepsis and septic shock: What does it give us? Med Intensiva 41:8–10 . doi: 10.1016/j.medin.2016.03.008

43. Segura-Sampedro J (2017) ¿Debemos asumir la nueva definición de sepsis en el campo de la cirugía? Cirugía Española

44. Vincent JL, Martin GS, Levy MM (2016) qSOFA does not replace SIRS in the definition of sepsis. Crit Care 20:1–3 . doi: 10.1186/s13054-016-1389-z

45. Simpson SQ (2016) New sepsis criteria A change we should not make. Chest 149:1117–1118 . doi: 10.1016/j.chest.2016.02.653

46. Peach BC (2017) Implications of the new sepsis definition on research and practice. J Crit Care 38:259–262 . doi: 10.1016/j.jcrc.2016.11.032

47. Angus DC, Linde-Zwirble WT, Lidicker J, et al (2001) Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Crit Care Med 29:1303–1310 . doi: 10.1097/00003246-200107000-00002

48. Martin GS, Mannino DM, Eaton S, Moss M (2003) The Epidemiology of Sepsis in the United States from 1979 through 2000. N Engl J Med 348:1546–1554 . doi: 10.1056/NEJMoa022139

49. Desautels T, Calvert J, Hoffman J, et al (2016) Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. JMIR Med Informatics 4:e28 . doi: 10.2196/medinform.5909

50. Slade E, Tamber PS, Vincent JL (2003) The surviving sepsis campaign: Raising awareness to reduce mortality. Crit Care 7:1–2 . doi: 10.1186/cc1876

51. Fleischmann C, Scherag A, Adhikari NKJ, et al (2016) Assessment of Global Incidence and Mortality of Hospital-treated Sepsis. 193:259–272 . doi: 10.1164/rccm.201504-0781OC

52. Walkey AJ, Lagu T, Lindenauer PK (2015) Trends in sepsis and infection sources in the United States. A population-based study. Ann Am Thorac Soc 12:216–220

53. Fleischmann C, Scherag A, Adhikari NKJ, et al (2016) Assessment of global incidence and mortality of hospital-treated sepsis. Current estimates and limitations. Am J Respir Crit Care Med 193:259–272

54. Rodríguez F, Barrera L, De La Rosa G, et al (2011) The epidemiology of sepsis in Colombia: A prospective multicenter cohort study in ten university hospitals. Crit Care Med 39:1675–1682 . doi: 10.1097/CCM.0b013e318218a35e

55. Ortiz G, Dueñas C, Rodrigez F, et al (2014) Epidemiology of sepsis in Colombian intensive care units. Biomédica 34: . doi: 10.1590/S0120-41572014000100007

56. Yende S, Angus DC (2007) Long-term outcomes from sepsis. Curr Infect Dis Rep 9:382–386

57. Weycker D, Akhras KS, Edelsberg J, et al (2003) Long-term mortality and medical care charges in patients with severe sepsis. Crit Care Med 31:2316–2323 . doi: 10.1097/01.CCM.0000085178.80226.0B

58. Iwashyna TJ, Ely EW, Smith DM, Langa KM (2010) Long-term cognitive impairment and functional disability among survivors of severe sepsis. JAMA - J Am Med Assoc 304:1787–1794 . doi: 10.1001/jama.2010.1553

59. Winters BD, Eberlein M, Leung J, et al (2010) Long-term mortality and quality of life in sepsis: a systematic review. Crit Care Med 38:1276–1283

60. Wang HE, Szychowski JM, Griffin R, et al (2014) Long-term mortality after communityacquired sepsis: A longitudinal population-based cohort study. BMJ Open 4:1–8 . doi: 10.1136/bmjopen-2013-004283

61. Ou SM, Chu H, Chao PW, et al (2016) Long-term mortality and major adverse cardiovascular events in sepsis survivors a nationwide population-based study. Am J Respir Crit Care Med 194:209–217 . doi: 10.1164/rccm.201510-2023OC

62. Arabi Y, Al Shirawi N, Memish Z, et al (2003) Assessment of six mortality prediction models in patients admitted with severe sepsis and septic shock to the intensive care unit: a prospective cohort study. Crit care 7:R116

63. Le Gall J-R, Lemeshow S, Leleu G, et al (1995) Customized probability models for early severe sepsis in adult intensive care patients. Jama 273:644–650

64. Carrara M, Baselli G, Ferrario M (2015) Mortality Prediction Model of Septic Shock Patients Based on Routinely Recorded Data. Comput Math Methods Med 2015:761435 . doi: 10.1155/2015/761435

65. Zhang Z, Hong Y (2017) Development of a novel score for the prediction of hospital mortality in patients with severe sepsis: the use of electronic healthcare records with LASSO regression. Oncotarget. doi: 10.18632/oncotarget.17870

66. Ford DW, Goodwin AJ, Simpson AN, et al (2016) A Severe Sepsis Mortality Prediction Model and

Score for Use With Administrative Data. Crit Care Med 44:319–327 . doi: 10.1097/CCM.0000000000001392

67. Lagu T, Lindenauer PK, Rothberg MB, et al (2011) Development and validation of a model that uses enhanced administrative data to predict mortality in patients with sepsis. Crit Care Med 39:2425–30 . doi: 10.1097/CCM.0b013e31822572e3

68. Osborn TM, Phillips G, Lemeshow S, et al (2014) Sepsis Severity Score. Crit Care Med 42:1969–1976 . doi: 10.1097/CCM.0000000000000416

69. Gall J-R Le (1995) Customized Probability Models for Early Severe Sepsis in Adult Intensive Care Patients. JAMA J Am Med Assoc 273:644 . doi: 10.1001/jama.1995.03520320054041

70. Winters BD, Eberlein M, Leung J, et al (2010) Long-term mortality and quality of life in sepsis: A systematic review. Crit Care Med 38:1276–1283 . doi: 10.1097/CCM.0b013e3181d8cc1d

71. Moody GB, Mark RG (1996) A database to support development and evaluation of intelligent\nintensive care monitoring. Comput Cardiol 1996 657–660 . doi: 10.1109/CIC.1996.542622

72. Lee J, Scott DJ, Villarroel M, et al (2011) Open-access MIMIC-II database for intensive care research. Proc Annu Int Conf IEEE Eng Med Biol Soc EMBS 8315–8318 . doi: 10.1109/IEMBS.2011.6092050

73. Saeed M, Lieu C, Raber G, Mark RG (2002) MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. Comput Cardiol 29:641–644 . doi: 10.1109/CIC.2002.1166854

74. Lehman L, Moody G, Heldt T, Kyaw TH (2011) NIH Public Access. Crit Care 39:952–960 . doi: 10.1097/CCM.0b013e31820a92c6.Multiparameter

75. Mikhno A, Ennett CM (2012) Prediction of Extubation Failure for Neonates with Respiratory Distress Syndrome Using the MIMIC - II Clinical Database. 2012 Annu Int Conf IEEE Eng Med Biol Soc IEEE,

76. Zhang Z, Xu X, Ni H, Deng H (2014) Predictive Value of Ionized Calcium in Critically Ill Patients : An Analysis of a Large Clinical Database MIMIC. PLoS One 9: . doi: 10.1371/journal.pone.0095204

77. Sun JX, Reisner AT, Saeed M, Mark RG (2005) Estimating Cardiac Output from Arterial Blood Pressure Waveforms : a Critical Evaluation using the MIMIC II Database. Comput Cardiol

78. Malhotra A, Waikar SS, Howell MD (2012) Outcome of Critically ill Patients with Acute Kidney Injury using the AKIN Criteria. Crit Care Med 39:2659–2664 . doi: 10.1097/CCM.0b013e3182281f1b.Outcome

79. Aboukhalil A, Nielsen L, Saeed M, et al (2008) Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. J Biomed Inform 41:442–451

80. Johnson AEW, Pollard TJ, Shen L, et al (2016) MIMIC-III, a freely accessible critical care database. Sci data 3: . doi: 10.1038/sdata.2016.35

81. Collins FS, Hudson KL, Briggs JP, Lauer MS (2014) PCORnet: turning a dream into reality. 576–577

82. M I T Critical Data (2016) Secondary Analysis of Electronic Health Records. Springer International Publishing

83. Pollard TJ, Johnson AEW, Raffa JD, et al (2018) The eICU collaborative research database, a freely available multi-center database for critical care research. Sci Data 5:1–13 . doi: 10.1038/sdata.2018.178

84. Cooke CR, Iwashyna TJ (2013) Using existing data to address important clinical questions in critical care. Crit Care Med 41:886–896 . doi: 10.1097/CCM.0b013e31827bfc3c

85. NIH The BioLINCC Handbook: A Guide to Accessing the NHLBI Biologic

86. Harrison D, Brady A, Rowan K (2004) Case mix, outcome and length of stay for admissions to adult, general critical care units in England, Wales and Northern Ireland: the Intensive Care National Audit & Research Centre Case Mix Programme Database. Crit Care 8:R99–R111 . doi: 10.1186/cc2834

87. Van Walraven C, Austin PC, Jennings A, et al (2009) A modification of the elixhauser comorbidity measures into a point system for hospital death using administrative data. Med Care 47:626–633 . doi: 10.1097/MLR.0b013e31819432e5

88. Elixhauser A, Steiner C, Harris DR, Coffey RM (1998) Comorbidity Measures for Use with Administrative Data. Med Care 36:8–27 . doi: 10.1097/00005650-199801000-00004

89. Shankar-Hari M, Rubenfeld GD (2016) Understanding Long-Term Outcomes Following Sepsis: Implications and Challenges. Curr Infect Dis Rep 18: . doi: 10.1007/s11908-016-0544-7

90. Winters BD, Eberlein M, Leung J, et al (2010) Long-term mortality and quality of life in sepsis: A systematic review*. Crit Care Med 38:1276–1283 . doi: 10.1097/CCM.0b013e3181d8cc1d

91. Yende S, Austin S, Rhodes A, et al (2016) Long-Term Quality of Life Among Survivors of Severe Sepsis. Crit Care Med 44:1461–1467 . doi: 10.1097/CCM.0000000000001658

92. Zhang Y, Haghani A (2015) A gradient boosting method to improve travel time prediction. Transp Res Part C 58:308–324 . doi: 10.1016/j.trc.2015.02.019

93. Jian Z, Shi X, Huang R, et al (2016) Feasibility of stochastic gradient boosting approach for predicting rockburst damage in burst-prone mines. Trans Nonferrous 26:1938–1945 . doi: 10.1016/S1003-6326(16)64312-1

94. Godinho S, Guiomar N, Gil A (2016) Using a stochastic gradient boosting algorithm to analyse the effectiveness of Landsat 8 data for montado land cover mapping: Application in southern Portugal. Int J Appl Earth Obs Geoinf 49:151–162 . doi: 10.1016/j.jag.2016.02.008

95. Brillante L, Gaiotti F, Lovat L, et al (2015) Investigating the use of gradient boosting machine, random forest and their ensemble to predict skin flavonoid content from berry physical-mechanical characteristics in wine grapes. Comput Electron Agric 117:186–193 . doi: 10.1016/j.compag.2015.07.017

96. Sut N, Simsek O (2011) Comparison of regression tree data mining methods for prediction of mortality in head injury. Expert Syst Appl 38:15534–15539 . doi: 10.1016/j.eswa.2011.06.006

97. Dietterich TG (2000) Ensemble Methods in Machine Learning. Mult Classif Syst 1857:1–15 . doi: 10.1007/3-540-45014-9

98. Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38:367–378

99. Friedman JH (2011) Greedy function machine: A gradient boosting machine. Statistics (Ber)

29:1189–1232 . doi: doi:10.1214/aos/1013203451

100. Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77:802–813 . doi: 10.1111/j.1365-2656.2008.01390.x

101. Kuhn M, Johnson K (2013) Applied Predictive Modeling. Springer

102. Tibshirani R (1996) Regression Shrinkage and Selection via the Lasso Robert Tibshirani. J R Stat Soc Ser B Stat Methodol 58:267–288 . doi: 10.1111/j.1467-9868.2011.00771.x

103. Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning. Math Intell 27:83–85 . doi: 10.1198/jasa.2004.s339

104. Vincent JL, Mendonça A De, Cantraine F, et al (1998) Use of the SOFA score to asses the incidence of organ dysfunction/failure in intensive care units: Results of a multicenter, prospective study. Crit Care Med 26:1793–1800

105. Vincent JL, Moreno R, Takala J, et al (1996) 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 Med 22:707–710 . doi: 10.1007/BF01709751

106. Johnson AEW, Kramer A a, Clifford GD (2013) A new severity of illness scale using a subset of Acute Physiology And Chronic Health Evaluation data elements shows comparable predictive accuracy. Crit Care Med 41:1711–8 . doi: 10.1097/CCM.0b013e31828a24fe

107. Paul P, Pennell ML, Lemeshow S (2013) Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. Stat Med 32:67–80 . doi: 10.1002/sim.5525

108. Kuhn AM, Wing J, Weston S, Williams A (2007) The caret Package. R Found Stat Comput

109. Ridgeway G (2007) Generalized Boosted Models : A guide to the gbm package. Compute 1:1–12 . doi: 10.1111/j.1467-9752.1996.tb00390.x

110. Jerome A, Hastie T, Simon N, Tibshirani R (2017) Package “ glmnet ”

111. Friedman JH, Meulman JJ (2003) Multiple additive regression trees with application in epidemiology. Stat Med 22:1365–1381 . doi: 10.1002/sim.1501

112. Legrand M, Payen D (2011) Understanding urine output in critically ill patients. Ann Intensive Care 1:13 . doi: 10.1186/2110-5820-1-13

113. Bermúdez-Rengifo W, Fonseca-Ruiz N (2016) Utilidad del lactato en el paciente críticamente enfermo. Acta Colomb Cuid Intensivo 15:13–18

114. Seleno N (2011) Elevation of blood urea nitrogen is predictive of long-term mortality in critically ill patients independent of “Normal” creatinine. J Emerg Med 40:724 . doi: 10.1016/j.jemermed.2011.04.013

115. Haq A, Patil S, Parcells AL, Chamberlain RS (2014) The Simplified Acute Physiology Score III is superior to the Simplified Acute Physiology Score II and Acute Physiology and Chronic Health Evaluation II in predicting surgical and ICU mortality in the “oldest old.” Curr Gerontol Geriatr Res 2014:

116. Hu Y, Zhang X, Liu Y, et al (2013) APACHE IV is superior to MELD scoring system in predicting

prognosis in patients after orthotopic liver transplantation. Clin Dev Immunol 2013:

117. Sharafoddini A, Dubin JA, Lee J (2017) Patient Similarity in Prediction Models Based on Health Data: A Scoping Review. JMIR Med Informatics 5:e7 . doi: 10.2196/medinform.6730

118. Poses RM, McClish DK, Smith WR, et al (1996) Prediction of survival of critically ill patients by admission comorbidity. J Clin Epidemiol 49:743–747

119. Sundararajan V, Henderson T, Perry C, et al (2004) New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. J Clin Epidemiol 57:1288–1294

120. Charlson M, Szatrowski TP, Peterson J, Gold J (1994) Validation of a combined comorbidity index. J Clin Epidemiol 47:1245–1251

121. Charlson ME, Pompei P, Ales KL, MacKenzie CR (1987) A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 40:373–383

122. Lee J, Maslove DM, Dubin JA (2015) Personalized mortality prediction driven by electronic medical data and a patient similarity metric. PLoS One 10:e0127428

123. Zhang Z, Mayer G, Dauvilliers Y, et al (2018) Exploring the clinical features of narcolepsy type 1 versus narcolepsy type 2 from European Narcolepsy Network database with machine learning. Sci Rep 8: . doi: 10.1038/s41598-018-28840-w

124. Schapire RE (1990) The strength of weak learnability. Mach Learn 5:197–227

125. Kuhn M, others (2008) Building predictive models in R using the caret package. J Stat Softw 28:1–26

126. Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorobot 7:21

127. Dignam JD, Martin PL, Shastry BS, Roeder RG (1983) Eukaryotic gene transcription with purified components. Methods Enzymol 101:582–598 . doi: 10.1016/0076-6879(83)01039-3

128. Meyer D, Buchta C (2019) Package “ proxy ”

129. Parisot S, Ktena SI, Ferrante E, et al (2017) Spectral graph convolutions for population-based disease prediction. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 10435 LNCS:177–185 . doi: 10.1007/978-3-319-66179-7_21

130. Kipf TN, Welling M (2016) Semi-Supervised Classification with Graph Convolutional Networks. arXiv Prepr 1–14 . doi: 10.1051/0004-6361/201527329

131. Zhu X, Ghahramani Z, Lafferty JD (2003) Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of the 20th International conference on Machine learning (ICML-03). pp 912–919

132. Yang Z, Cohen WW, Salakhutdinov R (2016) Revisiting Semi-Supervised Learning with Graph Embeddings. arXiv Prepr 160308861 48:

133. Kingma DP, Ba J (2014) Adam: A Method for Stochastic Optimization. arXiv Prepr 14126980 1–15

134. Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning. Springer series in statistics Springer, Berlin

135. Beeley C (2013) Web application development with R using Shiny. Packt Publishing Ltd