UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

MASTER IN BIOMEDICAL ENGINEERING

MASTER'S DISSERTATION

DATA ANALYSIS FOR THE EVOLUTION IN PATIENTS WITH TRAUMATISMS

ÁLVARO BELMAR MAS

2020

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

MASTER IN BIOMEDICAL ENGINEERING

MASTER'S DISSERTATION

DATA ANALYSIS FOR THE EVOLUTION IN PATIENTS WITH TRAUMATISMS

ÁLVARO BELMAR MAS

Tutor:

BLANCA LARRAGA GARCÍA

2020

Resumen

El trauma es la principal causa mundial de muerte entre las personas menores de 40 años. Además, el trauma severo es un importante problema de salud pública global, que contribuye a aproximadamente 1 de cada 10 muertes y resulta en la muerte mundial anual de más de 5,8 millones de personas. Por lo tanto, registrar información de todos estos casos ha sido una tarea muy importante en los últimos años.

Las características de los pacientes que sufrieron un traumatismo se recopilan en una base de datos junto con el estado de su estadía en el hospital y toda la información sobre los procedimientos realizados al paciente y las estancias en la Unidad de Cuidados Intensivos (UCI) si ese fuera el caso. Trabajaremos con la base de datos TQP PUF AY 2016, que contiene todos los registros enviados al Banco Nacional de Datos de Trauma (NTDB) hasta 2016.

Esta Tésis de Máster se enfoca en poder crear un modelo predictivo que pueda integrarse con simulaciones médicas, existentes o nuevas, para mejorar el entrenamiento impartido a los equipos de servicios médicos de emergencia (EMS) y del departamento de emergencia (ED). Y en analizar la base de datos de traumatismos TQP PUF de 2016, compuesta de 32 tablas con distintos tipos de datos (por ejemplo, demografía, signos vitales, alta de pacientes), tratando de encontrar correlaciones en los datos que nos permitan explotar aún más la relación para tratar de modelar un sistema de predicción. El proyecto se centraría en el estudio de la evolución de los pacientes con trauma en el hospital, un análisis profundo de los diferentes índices que existen (por ejemplo, Glasgow Coma Scale (GCS), Abbreviated Injury Scale (AIS)) y su papel en la definición del estado de un paciente con trauma.

Este objetivo se ha logrado parcialmente. Aunque se realizó una serie de operaciones para limpiar y preparar la base de datos, con el fin de mejorar los resultados del análisis de correlación, no se encontró una correlación significativa que pudiera integrarse en un sistema de pronóstico.

Sin embargo, el análisis de la base de datos se realizó con éxito, y pudimos conocer la precisión de los distintos índices de trauma evaluados: Glasgow Coma Scale (GCS), Injury Severity Score (ISS), Revised Trauma Score (RTS) y Trauma and Injury Severity Score (TRISS); y cómo los distintos parámetros involucrados en el trauma, como los signos vitales de los pacientes, y el tipo y la ubicación de la lesión, afectan al resultado del paciente.

Palabras clave: Análisis de datos, modelos de pronóstico, predicción, análisis de correlación, índices de trauma.

Abstract

Trauma is the main worldwide cause of death among people under 40 years old. Furthermore, severe trauma is a major global public health issue, contributing to about 1 in 10 mortalities and resulting in the annual worldwide death of more than 5,8 million people. Therefore, to register information of all these cases has been a very important task in the past years.

The characteristics of patients who suffered a traumatism are gathered in a database along with the status in their stay in the hospital and all the information regarding procedures done to the patient and Intensive Care Unit (ICU) stays if that were the case. We will be working with the database TQP PUF AY 2016, which contains all records sent to the National Trauma Data Bank (NTDB) up to 2016.

This Master Thesis focuses on being able to create a predictive model that could be integrated with existing or new medical simulations in order to improve the training given to the Emergency Medical Services (EMS) and Emergency Department (ED) teams. And to analyze the TQP PUF trauma database from 2016, composed of 32 tables with distinct types of data (e.g. demography, vitals, discharge), trying to find correlations in the data enabling us to further exploit the relation to try and model a prediction system. The project would focus on the study of the evolution of patients with trauma in the hospital, a deep analysis of the different indexes that exist (e.g. Glasgow Coma Scale (GCS), Abbreviated Injury Scale (AIS)) and their role at defining the state of a patient with trauma.

This objective has been achieved partially, although a series of operations to clean and prepare the database was performed, in order to improve the results of the correlation analysis, no significant correlation was found that could be integrated into a prognostic system.

However, the analysis of the database was successfully carried out, and we could learn of the accuracy of the distinct trauma scores evaluated: Glasgow Coma Scale (GCS), Injury Severity Score (ISS), Revised Trauma Score (RTS), and Trauma and Injury Severity Score(TRISS); and how the distinct parameters involved in trauma, such as, patients vital signs, and type and location of injury, affected the patient outcome.

Keywords: Data analysis, prognostic models, prediction, correlation analysis, trauma scores.

Acknowledgments

In the first place I would like to thank my tutor Blanca Larraga García for giving me the opportunity to work in this Master Thesis, and help and support me along the course of this work.

I would also like to thank my friends and family for being there and supporting me when the finish line didn't seem so clear, and of course, for listening to me ranting at all times when the stressful moments came.

Finally, a special thanks to my parents, that gave me the opportunity of coming to Madrid to do the master, and supported me in that decision.

Education is not the filling of a pail,
but the lighting of a fire.

William Butler Yeats.

Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

AIS	Abbreviated Injury Scale.
CPT	Current Procedural Terminology.
CSV	Comma-separated values.
CT	Computed Tomography.
ED	Emergency Department.
EHR	Electronic Health Record.
EMS	Emergency Medical Services.
GCS	Glasgow Coma Scale.
GPI	Generic Product Identifier.
ICD	International Classification of Diseases.
ICU	Intensive Care Unit.
ISS	Injury Severity Score.
LMICs	Low Middle Income Countries.
LOF	Local Outlier Factor.
NISS	New Injury Severity Score.
NTDB	National Trauma Data Bank.
RAM	Random-access memory.
RDBMS	Relational Database Management System.
RDSMS	Relational Data Stream Management System.
RR	Respiratory Rate.
RTS	Revised Trauma Score.
SBP	Systolic Blood Pressure.
SQL	Structured Query Language.
T-RTS	Revised Trauma Score - Triage.
TBI	Traumatic Brain Injury.
TRISS	Trauma and Injury Severity Score.
TS	Trauma Score.
TTM	Targeted Temperature Management.
US	United States.

1. Introduction

This Master Thesis aims to find and exploit correlations or relations between dependent variables to try and model, if possible, a way to predict the outcome of patients with trauma.

1.1. Thesis focus

Trauma is the main worldwide cause of death among people under 35 years old [1]. Furthermore, severe trauma is a major global public health issue, contributing to about 1 in 10 mortalities and resulting in the annual worldwide death of more than 5.8 millions people. The three main types of physical trauma are:

Blunt force trauma: When an object or force strikes the body.

Penetrating trauma: When an object pierces the skin or the body.

Burn trauma: The result of a burn in an area of the body.

Therefore, to register information of all these cases has been a very important task in the past years.

In the United States the characteristics of patients who suffered a traumatism are gathered in a database along with the status in their stay in the hospital and all the information regarding the procedures done to the patient and the Intensive Care Unit (ICU) stays if that were the case. We will be working with the TQP PUF AY 2016 database, which contains all records sent to the National Trauma Data Bank (NTDB) up to 2016 [2].

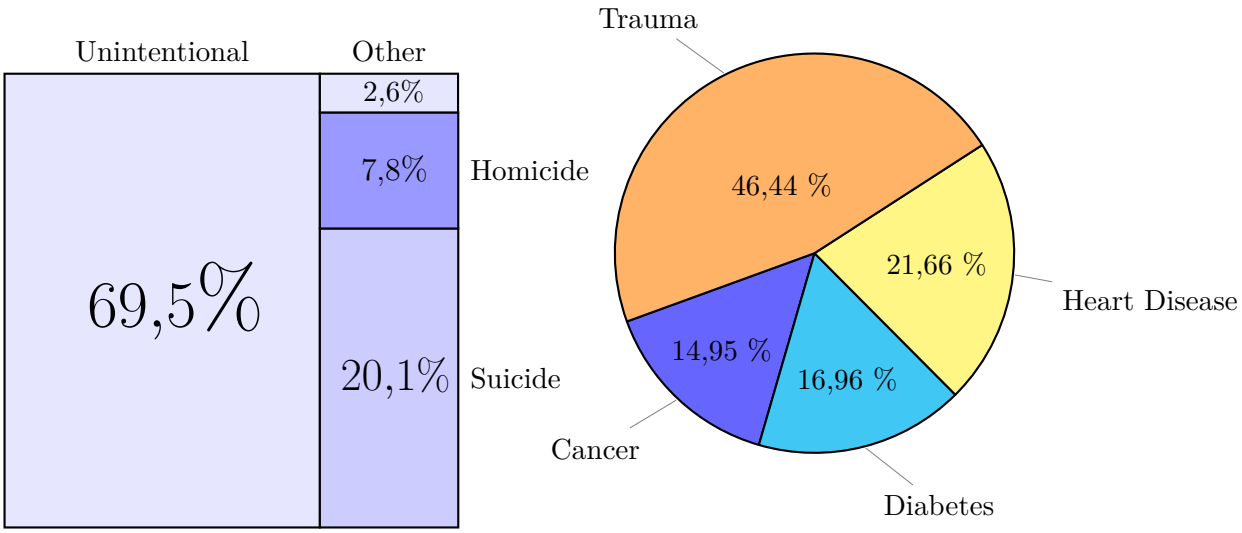
This data will be analyzed to try and find correlations between the distinct types of variables composing the database. Therefore, finding a way to predict the outcome of a patient, or to better treat a patient based on the trauma suffered supporting the procedure using verified data.

The final objective of this Master Thesis would be to be able to create a predictive model that could be integrated with existing or new medical simulations in order to improve the training given to the Emergency Medical Services (EMS) and Emergency Department (ED) teams.

1.2. Trauma

Traumatic injury is a term which refers to physical injuries of sudden onset and severity which require immediate medical attention to treat the patient.

Approximately 90% of these injury deaths occur in Low Middle Income Countries (LMICs) [3]. As trauma affects to relatively young population, produces a higher economical and social impact than other illnesses [4].



(a) Causes of injury related death

(b) Percentage of the annual cost of trauma, total = $1.445 B

Figure 1.1.: Trauma statistics in the US [5]

In Figure 1.1 the cost of the trauma disease per year is shown. This amount rises to $671 B being $1.445 B the total cost in healthcare services, representing almost the 50% of the costs. Furthermore, it is also worth noting that although 69,5% of deaths are unintentional (e.g. car crashes, falls), 20,1% are related to suicide cases [5], depending on the country this number can vary.

In Spain, in 2018, 15.768 trauma related deaths were registered, of those, 3.539 (22,4%) were suicides and self-inflicted injuries [6].

As the three main mechanisms of trauma are: penetrating, blunt, and burns. Studies [7, 8] have found that blunt traumas are the most dangerous causing several injuries to patients and increasing the hospital, intensive care unit, and mechanical ventilation days, while burn traumas are the less common, with only 0,2 to 2,9/10.000 people.

The most common types of traumatic injuries [9] are:

- Traumatic brain injury
- Spinal cord injury
- Spine fractures
- Amputation - traumatic
- Facial trauma

- Acoustic trauma
- Crush injury
- Concussion
- Broken bone
- Jaw - Broken or dislocated
- Skull fracture
- Cuts and puncture wounds
- Collapsed lung
- Myocardial contusion
- Burns
- Electrical injury
- Hypovolemic shock
- Subarachnoid hemorrhage
- Subdural hematoma

Trauma deaths occur in immediate, early, or late stages according to the trimodal distribution of trauma deaths [10]. **Immediate deaths** usually are due to apnea, severe brain injury, or rupture of the heart or of large blood vessels, and they occur at the same time as the injury happens. **Early deaths** occur within minutes to hours and often are due to hemorrhages in the outer meningeal layer of the brain, torn arteries, blood around the lungs (hemothorax), air around the lungs (pneumothorax), ruptured spleen, liver laceration, or pelvic fracture. **Late deaths** occur days or weeks after the injury and are often related to infections [11].

In the case of immediate deaths, these are nonsurvivable injuries in which only preventing the injury in the first place can affect the outcome for the patient, immediate deaths represent the 50% of the total trauma deaths. For early deaths, the access to a care facility is important; this group represents the 30% of the total deaths caused by trauma. Lastly, late deaths can be prevented with improved resuscitation techniques and critical care, this group represents the 20% of trauma deaths [12].

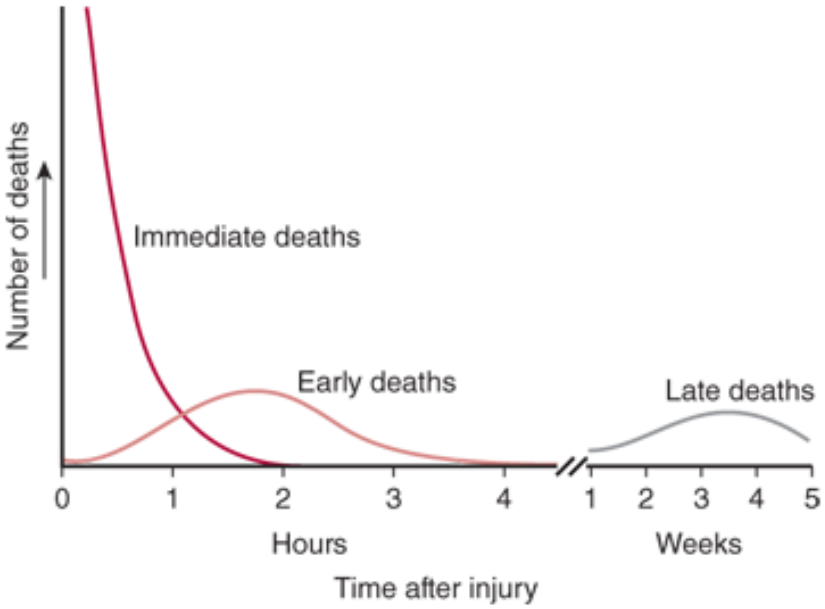


Figure 1.2.: Number of deaths by time elapsed from trauma [12]

Among all the possible causes of death for trauma patients, bleeding and Traumatic Brain Injury (TBI) carried the greater risk of death [13]. Patients with TBI carried about 2.5 times the risk of death compared with the general population. Life expectancy reduction averaged 6 years [14].

In the case of post-traumatic bleeding, uncontrolled post-traumatic bleeding is the leading cause of potentially preventable death among trauma patients [15]. About one-third of all trauma patients with bleeding present with a coagulopathy on hospital admission [16]. This subset of patients has a significantly increased incidence of multiple organ failure and death compared to patients with similar injury patterns in the absence of a coagulopathy [17]. Moreover, hemorrhagic shock occurs when the body begins to shut down due to large amounts of blood loss. Hemorrhagic shock represents a 35.2% of trauma deaths [18].

1.3. Clinical simulation

Clinical simulation consists on a series of programs and practical seminaries in which medical students, ED personnel, and EMS personnel, learn to approach real life intervention situations using systems created to simulate those situations (virtual programs, mannequins). This practice helps gain the experience needed to perfectly perform the procedures without hesitation nor error.

The importance of medical simulation comes from studies that show that human errors constitutes a high number of deaths and medicinal costs [19]. Medical errors

cause injuries to approximately 3% of hospital patients, resulting in a number of deaths per year that could range from 44.000 to 98.000 patients in the United States [20].

If we were to be successful in finding strong correlations to exploit in order to create a predictive model, we could integrate those model along with medical simulations programs, which would in turn improve the skill of the teams in charge of major trauma events.

1.4. Document layout

This manuscript has been structured in four different sections.

Chapter 2: In this section we are going to talk about the state of the art of the area in which this project is focused, that is data analysis. Moreover, the techniques that we are going to use throughout the thesis will be presented. Within this section, the trauma scores and their uses in diagnosing the severity of the trauma are explained.

Chapter 3: This section explains the process followed during the Master Thesis development. The tools used in the thesis to process and analyze the data will be presented. Moreover, we will expose the structure of the database, so that the different tables and relations that we will use in the development and experimentation sections can be easily understood.

Chapter 4: In this section we will show an analysis of the data, how the different variables vary in function of the trauma type and the previous conditions of the patient. Furthermore, we will explain the process followed in order to calculate the correlations between tables of the database.

Chapter 5: This section will provide with the results of the process previously explained in the development section, as well as the conclusions of the correlation study.

2. Literature review

In this Chapter we are going to expose the state of the art of the trauma scores used in hospitals globally, as well as the correlation and prediction systems for clinical data.

2.1. Trauma scores

This section would be focused on reviewing the trauma scores used globally by hospitals as a pseudo-prognostic system able to determine the state and evolution of the patient.

2.1.1. Glasgow Coma Scale (GCS)

This scale was published in 1974 by Graham Teasdale and Bryan J. Jennett [21], it is a neurological scale which aims to give a reliable and objective way of recording the state of a person consciousness for initial as well as subsequent assessment. The resulting points give a person's score between 3 (indicating deep unconsciousness) and either 14 (original scale) or 15 (more widely used, modified or revised scale).

Glasgow Coma Scale (GCS) was used to assess a person's level of consciousness after a head injury, and the scale is now used by the EMS, nurses, and physicians to all acute medical and trauma patients. In hospitals, it is also used in monitoring patients in intensive care units [22].

GCS performance as a prognostic tool has been researched in several studies [23, 24, 25, 26]. One of them [25], found that GCS predicted correctly 26 deaths out of 104 patients of Traumatic Brain Injury (TBI), a 25% of accuracy. Another study [26], reached the conclusion that GCS values lower than 6,5 increased the mortality of patients significantly.

The value of this score is calculated as the sum of three types of responses: the ocular, the verbal and the motor responses. For any of the responses 3 is the lowest possible value and 15 is the highest one. A value higher or equal than 13 is considered a minor injury, a value between 9 and 12, is considered moderate, and lower than 9, is considered a severe injury. The three possible responses are as follows:

Ocular response: 1 to 4, 1 being unresponsive, 4 being responsive.

Verbal response: 1 to 5, 1 being none, 5 being oriented.

Motor response: 1 to 6, 1 being no results, 6 being obeys commands.

2.1.2. Abbreviated Injury Scale (AIS)

Abbreviated Injury Scale (AIS) is an anatomical-based coding system created by the Association for the Advancement of Automotive Medicine to classify and describe the severity of injuries [27]. It represents the threat to life associated with the injury, rather than the comprehensive assessment of the severity of the injury. It has several versions, with the first one being in 1969 and the last one in 2015.

The Abbreviated Injury Scale (AIS) classifies individual injuries by body region as follows:

- Head
- Face
- Neck
- Thorax
- Abdomen
- Spine
- Upper Extremity
- Lower Extremity
- External and other

Each of the body parts can receive one of the following scores:

- AIS 1 – Minor
- AIS 2 – Moderate
- AIS 3 – Serious
- AIS 4 – Severe
- AIS 5 – Critical
- AIS 6 – Maximal (currently untreatable)

2.1.3. Injury Severity Score (ISS)

The Injury Severity Score (ISS) is an established medical score to assess trauma severity [28]. It correlates with mortality, morbidity and hospitalization time after trauma. It is used to define the term major trauma.

To calculate the Injury Severity Score (ISS) of an injured person, the body is divided into six ISS body regions. The highest AIS in each of the three most injured ISS body

regions is taken to be used in Equation 2.1, where A, B, C are the AIS scores of the three most injured ISS body regions. The ISS scores ranges from 3 to 75 (low severity to maximum severity). If any of the three scores is a 6, the score is automatically set at 75.

As ISS takes the three highest values without focusing on which parts of the body are affected, it gives all body parts the same importance for the patient survival. An ISS score greater than 15 is considered a polytrauma [29].

The biggest disadvantage of the ISS score is the inability to take into account a specific area of the body, because of that it's only better than AIS at predicting accidental traumas, whereas AIS is better at predicting outcome of head injuries [30].

$$ISS = A^2 + B^2 + C^2 \tag{2.1}$$

The body regions are as follow:

1. Head or neck – including spine
2. Face
3. Chest
4. Abdomen or pelvic contents
5. Extremities or pelvic girdle
6. External

2.1.4. New Injury Severity Score (NISS)

The New Injury Severity Score (NISS) fixes the highest problem that ISS has, taking into account severe injuries that appear in a single body area. New Injury Severity Score (NISS) inputs into Equation 2.1 the three highest scores regardless of the anatomic area. Overall the accuracy of NISS was better than ISS, although the difference is more significant when the patient suffers a head/neck injury [31]. However, other studies show that the difference between NISS and ISS is not that significant without focusing on specific trauma types [32], as when taking the whole body into account both scores are essentially the same.

2.1.5. Trauma Score (TS)

The Trauma Score (TS) [33] measures the acute component of trauma taking into five variables. The sum of the points given to those five variables (Table 2.1) results in a number between 1 and 16, being 1 the worst possible result and 16 the best possible result.

The five variables are:

- GCS
- Respiratory Rate (RR)
- Respiratory effort
- Systolic Blood Pressure (SBP)
- Capillary refill

GCS	Points	SBP [mm Hg]	Points	RR [breath-s/min]	Points	Respiratory effort	Points	Capillary refill	Points
14-15	5	90	4	10-24	4	Normal	1	Normal	2
11-13	4	70-90	3	25-35	3	Shallow or retractive	0	Delayed	1
8-10	3	50-69	2	35	2			None	0
5-7	2	1-49	1	10	1				
3-4	1	0	0	0	0				

Table 2.1.: Scoring of variables in TS

As respiratory expansion and capillary refill were too hard to accurately measure RTS was developed.

2.1.6. Revised Trauma Score (RTS)

The Revised Trauma Score (RTS) [34] takes into account only three of the five variables that the TS considers: GCS, RR, and SBP. In Table 2.2 we can see the values appointed to each value in each category. A weighted sum is performed with those scores (Equation 2.2). As it can be seen, the score is heavily weighted towards GCS to account to the possibility of severe head injury. In the case of a RTS of less than 4, the survival rate drops to 50%. Nowadays, the RTS score is used together with the ISS score in order to achieve better prediction accuracy.

$$RTS = 0.9368 \cdot GCS + 0.7326 \cdot SBP + 0.2908 \cdot RR \tag{2.2}$$

2.1.7. Revised Trauma Score - Triage (T-RTS)

The Revised Trauma Score - Triage (T-RTS) is a version of RTS that is used in triage applications. It is the sum of the three RTS variables, ranging from, 0 most severe, to 12, least severe trauma lesion.

Triage is a tool used to classify people by the level of urgency in their condition and the likelihood of recovery without medical care. It is amply used by the EMS

Glasgow Coma Scale	Systolic Blood Pressure [mm Hg]	Respiratory Rate [breaths/min]	Points
15-13	>89	10-29	4
12-9	76-89	>29	3
8-6	50-75	6-9	2
5-4	1-49	1-5	1
3	0	0	0

Table 2.2.: Scoring of variables in RTS

services to adapt their choices regarding the urgency of the trauma [34, 35]. There are five levels:

- Red - 0 minutes to care, resuscitation
- Orange - 15 minutes to care, urgent
- Yellow - 60 minutes to care, less urgent
- Green - 120 minutes to care, standard
- Blue - 240 minutes to care, not urgent

2.1.8. Trauma and Injury Severity Score (TRISS)

The Trauma and Injury Severity Score (TRISS) is a method to calculate the probability of survival of a patient [36]. It uses RTS, ISS, and the patient age to calculate the coefficient b (Trauma and Injury Severity Score (TRISS)). Studies [37, 38] show that TRISS outperforms ISS and RTS in mortality prediction. Equation 2.3 shows the formula.

$$b = b_0 + b_1 \cdot RTS + b_2 \cdot ISS + b_3 \cdot A \tag{2.3}$$

where:

b_0, b_1, b_2, b_3 = weighted coefficients, see Table 2.3
A = constant, 1 if patient age is greater than 54, 0 if not

Using the coefficient from Equation 2.3, in Equation 2.4 the probability of survival of a trauma patient is obtained applying Equation 2.4 as follows:

$$Ps = \frac{1}{1 + e^{-b}} \tag{2.4}$$

2.2. Methods used in the literature

In this subsection we are going to explain the methods used for analyzing data.

	b_0	b_1	b_2	b_3
Blunt	-1,6465	0,5175	-0,0739	-1,9261
Penetrating	-0,8068	0,5442	-0,1159	-2,4782

Table 2.3.: Values for weighted coefficients in TRISS

2.2.1. Correlation

In statistics, correlation is a statistical relation, which can be casual or not, between variables that appear random. In general, it defines a statistical relation, although it is more commonly used to define the degree of linear relation of the two variables [39].

The importance of correlation comes from the fact that it can indicate a predictive relationship between two variables. This allows to exploit correlations in predictive systems or models [40]. Random variables are dependent if they do not satisfy the mathematical property of probabilistic independence, which can lead to correlation being used as a synonym of dependence [41].

There are several correlation coefficients, often denoted ρ, r, or τ, which measure the degree of correlation. These correlations coefficients have been widely used in medical research throughout the years [42, 43, 44, 45].

Pearson correlation coefficient

Pearson correlation coefficient is a statistic method that measures the linear correlation of two variables between -1 and +1, where 1 is total positive linear correlation, 0 means that there is no linear correlation, and -1 means that there is a total negative linear correlation [46], an example of these values and their representation can be seen in Figure 2.1.

The Pearson correlation coefficient is symmetric. That phenomena can be seen in the equation for Pearson's correlation coefficient (Equation 2.5) as $cov(X, Y) = cov(Y, X)$.

$$\rho_{X,Y} = \frac{cov(X, Y)}{\sigma_X \sigma Y} \tag{2.5}$$

where:

cov = covariance
σ_X = standard deviation of X
σ_Y = standard deviation of Y

Spearman's rank correlation coefficient

[1] Extracted from `https://en.wikipedia.org/wiki/Pearson_correlation_coefficient`

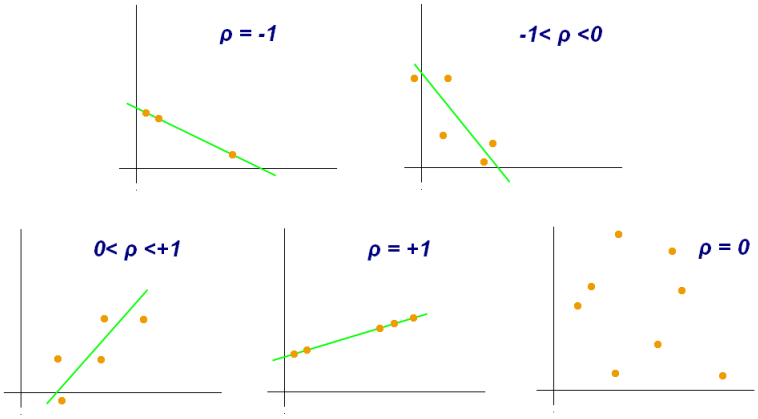
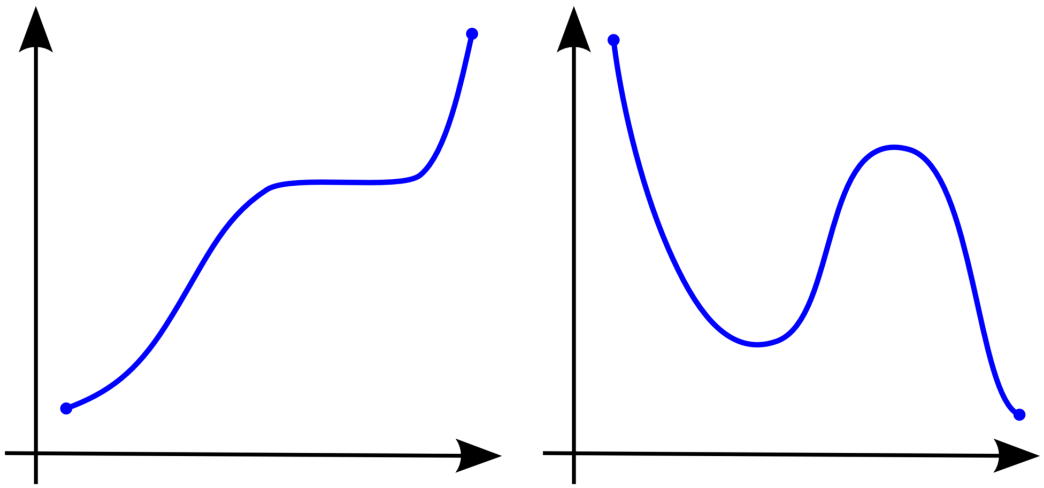


Figure 2.1.: Examples of scatter diagrams with different values of Pearson correlation coefficient ρ [1]

Spearman's rank correlation coefficient is a non-parametric measure of rank correlation, denoting the dependence of the rankings of two variables, assessing how accurate can be the description of the relation of two variables using a monotonic function (Figure 2.2).



(a) Monotonic function **(b)** Non-monotonic function

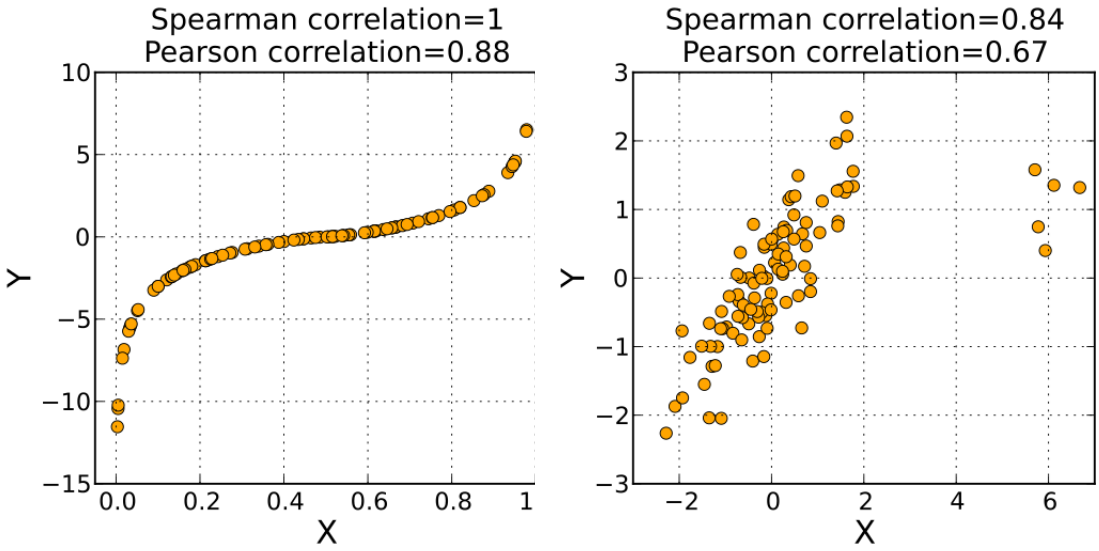
Figure 2.2.: Monotonic and non-monotonic functions [2]

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those variables. While Pearson assesses exclusively linear relationships, Spearman's correlation assesses monotonic relationships, which can be linear or not. This effect can be seen in Figure 2.3, where in 2.3a performs better than Pearson's correlation, where a Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. However, in 2.3b we can observe that Spearman is less sensitive than Pearson

[2] Extracted from `https://en.wikipedia.org/wiki/Monotonic_function`

to strong outliers, that is because Spearman's ρ limits the outlier to the value of its rank. In order to use this method of correlation, a thorough cleaning of the data set is mandatory.

Spearman's coefficient is appropriate for both continuous and discrete ordinal variables [47].



(a) In monotonic relationships Spearman's correlation performs better than Pearson's

(b) With strong outliers present Spearman's correlation performs worse than Pearson's

Figure 2.3.: Comparison between Spearman's correlation and Pearson's [3]

Kendall rank correlation coefficient

The Kendall rank correlation coefficient is used to measure the rank correlation between two measured quantities. A τ test is a non-parametric hypothesis test for dependence based on the Kendall's τ coefficient. It is a measure of rank correlation. Both Spearman's ρ and Kendall's τ can be formulated as special cases of a more general correlation coefficient [48].

The definition of Kendall's tau is the notion of concordance. If (x_i, y_i) and (x_j, y_j) are two elements of a sample from a bivariate population, one says that (x_i, y_i) and (x_j, y_j) are concordant if $x_i > x_j$ and $y_i > y_j$ or if $x_i < x_j$ and $y_i < y_j$; and discordant if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither discordant nor concordant.

The Kendall coefficient is defined as seen in Equation 2.6.

$$\tau = \frac{(number\,of\,concordant\,pairs) - (number\,of\,discordant\,pairs)}{\binom{n}{2}} \tag{2.6}$$

[3] Extracted from `https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient`

where:

$\binom{n}{2} = \frac{n(n-1)}{2}$ number of ways to choose two items from n items.

Moreover, Kraemer [49] shows a comparison between the coefficients, explaining in which situation they would perform better. When X and Y are ordinal variables, that is when they have order, and follow a bivariate normal distribution, Pearson's coefficient works best. However, if the assumption of those variables following a bivariate normal distribution were to be not fully asserted, Kendall's and Spearman's coefficients would outperform Pearson's coefficient in this situation. Moreover, Pearson is most widely used of the three.

2.2.2. Outliers

In statistics, an outlier is a data point that differs significantly from other observations [50]. Outliers, by being the most extreme observations of the data set, may include the maximum or minimum, or even both. However, the maximum or minimum are not always outliers because they may not be too far from other observations of the data set.

Estimators capable of dealing with outliers are said to be robust, for example, the median is a robust statistic variable of central tendency, while the mean is not [51]. However, the mean is generally a more precise estimator [52].

For the detection of the outliers, there is no rigid mathematical definition of what defines an outlier. Determining if an observation is an outlier or not dependes on the point of view [53]. Some methods of outlier detection include graphical forms, such as normal probability plots (Figure 2.4) where ouliers can be detected as strays out of the straight line, while others are model based [54, 55]. Box plots are a mix of the graphical and model based (Figure 2.5).

The two methods described below are both statistical methods used in other studies with the aim to remove outliers of the Electronic Health Record (EHR) [56, 57].

[4] Extracted from `https://en.wikipedia.org/wiki/Normal_probability_plot`

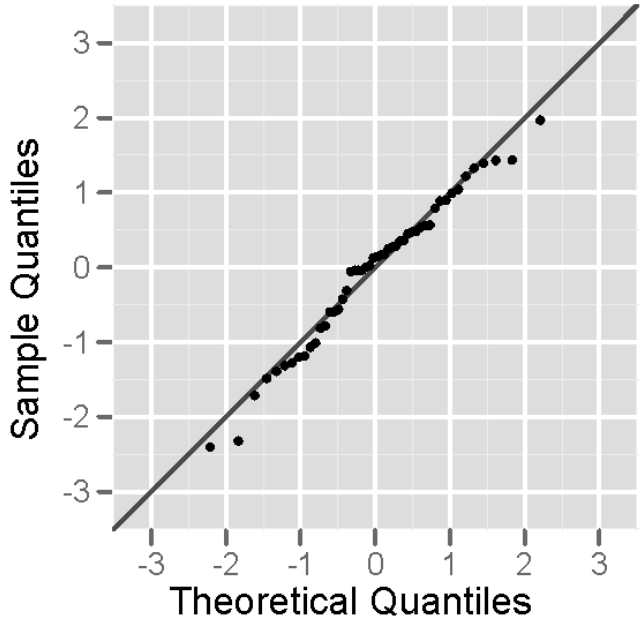


Figure 2.4.: Normal probability plot [4]

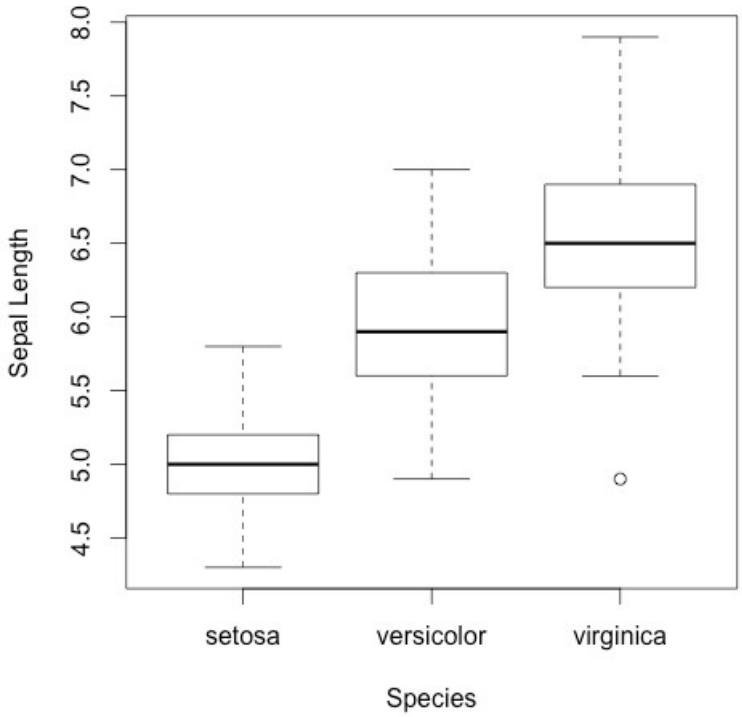


Figure 2.5.: Box plot of sepal length in the famous Iris[5]data set, in virginica species an outlier can be seen as a dot [6]

[6] `https://archive.ics.uci.edu/ml/datasets/iris`
[6] Extracted from `https://www.r-bloggers.com/use-box-plots-to-assess-the-distribution-and-to-identify-the-outl`

Tukey's fences

Tukey's fences method is based on the measure of the interquartile range. If Q_1 and Q_3 are the lower and upper quartiles respectively, an outlier can be defined as an observation outside the range as shown in Equation 2.7 [58].

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)] \tag{2.7}$$

where:

k = negative constant, Jhon Tukey proposed, k = 1.5 is an outlier, k = 3 is data "far out"

Anomaly detection

In data mining, anomaly detection focuses on the identification of rare items, events or observations which differ significantly from the rest of the data. This task is normally resolved using distance-based [59] and density-based models, such as Local Outlier Factor (LOF) [60], and the majority of them uses k-nearest neighbors to label an observation as outlier or not [61]. In Figure 2.6 we can observe how the outliers have a much lower density than other observations, the more isolated the observation is the lower density it will have. As observations become more "crowded" their relative space with other observations is lower, resulting in higher densities.

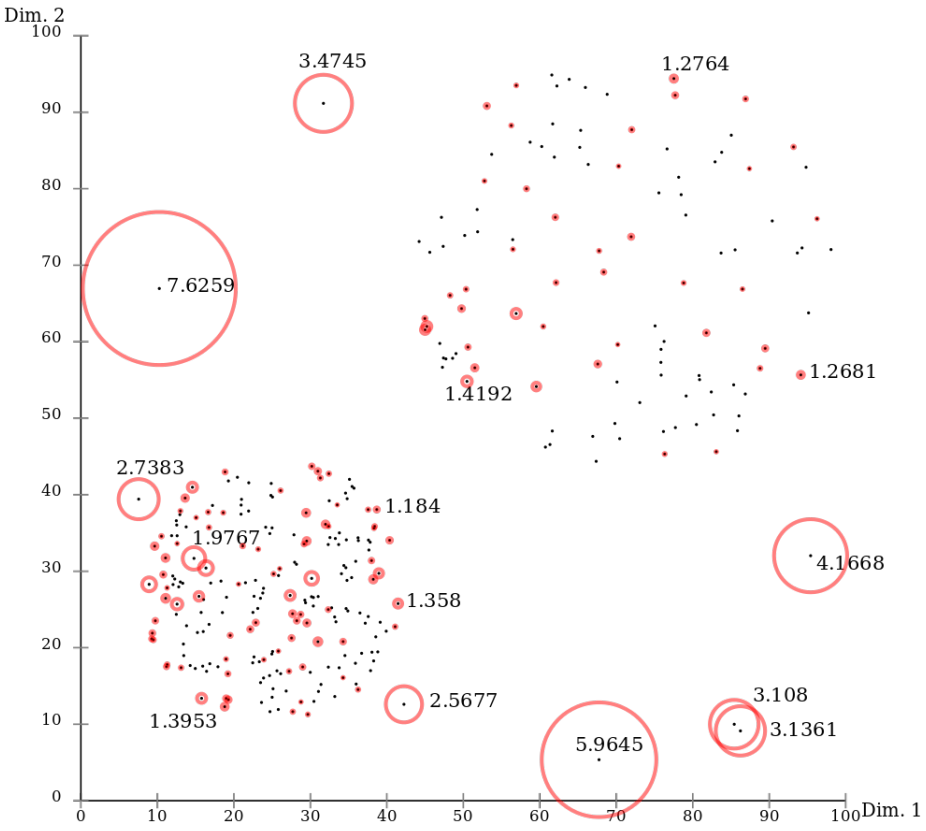


Figure 2.6.: Local Outlier Factor [7]

Z-score

The standard score, or Z-score, is the number of standard deviations by which the value of a data point is above or below the mean of the data set. The standardizing process, as it is called, is calculated by subtracting the population mean from an individual data point, and then dividing the difference by the population standard deviation [62].

When calculating the Z-score there can be two possibilities: the mean and the standard deviation of the population is known (Equation 2.8), or both are unknown, in which case the Z-score can be calculated using the sample mean and the sample standard deviation as an estimation (Equation 2.9).

$$z = \frac{x - \nu}{\sigma} \tag{2.8}$$

where:

x = data point
μ = mean of the data set
σ = standard deviation of the population

$$z = \frac{x - \bar{x}}{S} \tag{2.9}$$

where:

x = data point
$\bar{x}$ = mean of the sample
S = standard deviation of sample

Z-score defines an outlier when its value distances a lot from the mean, usually the values taken are +/-3 from the mean. The reason for this threshold can be seen in Figure 2.7, where 99.7% of the data is contained between +/-3 Z-score.

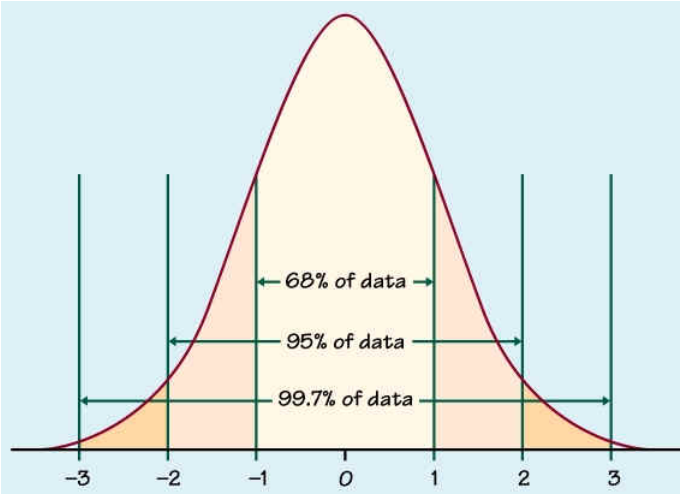


Figure 2.7.: Z-score distribution [8]

[7] Extracted from `https://en.wikipedia.org/wiki/Local_outlier_factor`
[8] Extracted from `http://www.ltcconline.net/greenl/courses/201/probdist/zScore.htm`

3. Methodology

This Chapter aims to detail the project development process, explaining the steps followed and the tools used throughout this Master Thesis.

Firstly we are going to see the structure of the trauma database TQP PUF AY 2016. This database contains the records of trauma patients in the US between the years 2007 and 2016. The importance of having a national trauma database is significant, as it can help pushing medical research forward, thus indirectly affecting lives and optimizing medical work flows.

3.1. Database structure

TQP PUF AY 2016 consists of 32 tables, each one with its own type of information, and related to others tables by a column key called *INC_KEY* that serves as an identifier of the patient. The whole structure can be seen in Table 3.1.

The first six tables, the ones starting with the prefix AIS provide information about the Abbreviated Injury Scale. Here we can also see the division that was created from the change from AIS98 to AIS05. The difference between the two codes resides primarily in the permissive actions that in 98 permitted others than physicians to report the grade of an injury, whereas in 05 version physicians needed to give approval [63]. The table PUF_AIS05TO98_CROSSWALK can be used to adapt the jump.

The database contains data of the demography of patients (PUF_DEMO), and their comorbidities (PUF_COMORBID). Moreover, it contains information about their stay at the hospital, such as the ED information (PUF_ED), the discharge information (PUF_DISCHARGE), the method of transport (PUF_TRANSPORT), and the complications that may have arisen (PUF_COMPLIC).

Furthermore, the database contains data about the actions done in the hospital, as well as the state of the patient. These tables collect the external cause of injury (PUF_ECODE, PUF_ICD10_ECODE), the diagnosis code (PUF_DCODE, PUF_ICD10_DCODE), the injury location codes (PUF_ICD10_LOC), procedure codes (PUF_PCODE), and patient vital signals while in ED and EMS (PUF_VITALS). The external cause of injury and the diagnosis codes tables have two versions, the ICD-9 and the ICD-10, ICD-10 contains a greater number of codes [64], the version used in the table is specified. The injury location code only has ICD-10 data, and PUF_PCODE collects both ICD versions, which are specified by a column inside the table. Moreover, all the codes in the database have their corresponding description tables, with the suffix DES, which convert the numerical codes to their descriptions. As PUF_PCODE has both versions, their description is stored in two different tables, PUF_ICD10_PROCDES for the ICD-10 version, and PUF_PCODEDES for the ICD-9 version.

Table 3.1.: Tables and descriptions included in TQP PUF AY 2016.

Begin of Table		
Table Name	Admission Years	Description
PUF_AIS05TO98_CROSSWALK	2016	Crosswalk from AIS 05 codes to AIS 98 codes
PUF_AIS98PCODE	2009 - 2015	The AIS code globally represented (submitted or mapped) to AIS version 1998. This table is no longer provided in 2016 however AIS 05 codes can be mapped to AIS 98 codes using the PUF_AIS05TO98_CROSSWALK table to crosswalk AIS 05 codes to AIS 98 codes.
PUF_AISCCODE	2007 - 2015	The AIS code globally calculated from ICD-9 diagnosis codes.
PUF_AISDES	2007 - 2016	AIS injury descriptors. AIS 98 exclusively until 2015 and then AIS 05 exclusively in 2016.
PUF_AISP05CODE	2007 - 2015	The AIS version 2005 code as submitted by the hospital. Note that this dataset does not contain AIS descriptors for AIS05 until 2016.
PUF_AISPCODE	2007 - 2016	The AIS (Abbreviated Injury Scale) code submitted by the hospital.
PUF_COMORBID	2007 - 2016	Comorbid conditions
PUF_COMPLIC	2007 - 2016	Hospital complications
PUF_DEMO	2007 - 2016	Demographic information
PUF_DCODE	2007 - 2016	ICD-9-CM diagnosis codes
PUF_DCODEDES	2007 - 2016	Lookup table of the description of the ICD-9-CM diagnosis codes
PUF_DISCHARGE	2007 - 2016	Includes discharge and outcome information
PUF_ECODE	2007 - 2016	Includes the ICD-9-CM external cause of injury code
PUF_ECODEDES	2007 - 2016	Lookup table of the description of the ICD-9-CM E-Codes
PUF_ED	2007 - 2016	Emergency department information
PUF_ICD10_DCODE	2015 - 2016	ICD-10-CM diagnosis codes
PUF_ICD10_DCODEDES	2015 - 2016	Lookup table of the description of the ICD-10-CM diagnosis codes

Continuation of Table 3.1

Table Name	Admission Years	Description
PUF_ICD10_ECODE	2015 - 2016	Includes the ICD-10-CM external cause of injury codes
PUF_ICD10_ECODEDES	2015 - 2016	Lookup table of the description of the ICD-10-CM E-Codes
PUF_ICD10_LOC	2015 - 2016	ICD-10-CM injury location codes
PUF_ICD10_LOCDES	2015 - 2016	Lookup table of the description of the ICD-10-CM location codes
PUF_ICD10_PROCDES	2015 - 2016	Lookup table of the description of the ICD-10-CM procedure codes
PUF_PCODE	2007 - 2016	ICD-9-CM and ICD-10-CM procedure codes (Pre-2015: ICD-9-CM only)
PUF_PCODEDES	2007 - 2016	Lookup table of the description of the ICD-9-CM procedure codes
PUF_PM	2013 - 2016	Information about the TQIP Processes of Care Measures elements. These elements are required from Level I and II TQIP centers only.
PUF_PM_EMBOLIZE_SITE	2013 - 2016	Information about the TQIP Processes of Care Measure element for embolization site. This element is required from Level I and II TQIP centers only.
PUF_PM_TBI_CM	2013 - 2016	Information about the TQIP Processes of Care Measure element for cerebral monitor. This element is required from Level I and II TQIP centers only.
PUF_PM_TBI_GCS_Q	2013 - 2016	Information about the TQIP Processes of Care Measure element for GCS assessment qualifiers. This element is required from Level I and II TQIP centers only.
PUF_PROTDEV	2007 - 2016	Protective devices
PUF_TRANSPORT	2007 - 2016	Transport information
PUF_VITALS	2007 - 2016	Vital signs from EMS and ED
TQP_INCLUSION	2010 - 2016	Information about a record's affiliation with a Trauma Quality Improvement Program (TQIP) institution, and whether that incident also met TQIP inclusion criteria for any of our reporting products

End of Table

A sample of a table (PUF_VITALS) can be seen in Table 3.2. In which the SBP, pulse, RR, oxygen saturation, temperature, and the sum of GCS are stored.

INC_KEY	SBP	PULSE	RR	OXYSAT	TEMP	GCSTOT
160967858	100	90	18	100	36.8	15
160967863	155	58	16	97	36.1	15
160759229	147	86	18	99	37.7	15

Table 3.2.: Sample of PUF_VITALS table contents

3.2. Analysis procedure

To achieve our objective we will analyze the database containing up to 32 tables with distinct types of data (e.g. demography, vitals, discharge), trying to find correlations in the data enabling us to further exploit the relation between variables to try and model a prognostic system.

The sequence of steps followed in the development process is as follows:

1. **Study of the state of the art in data analysis in the healthcare industry:**
   We performed and studied the state of the art in clinical data analysis, focusing on ways to effectively analyze the data, and on prognostic systems and their achieved results (Section 2.1).

2. **Study of the structure of the database TQP PUF year 2016:**
   In this step, we studied the structure of the database, the tables within it, and the type of data. Furthermore, we learned how to join the tables that were separated in two, the code and the code description tables, such as, the diagnosis and procedure codes tables. To do so, we used the reference codes as foreign keys in the relation.

3. **Cleanup of the database (e.g. outliers, missing data, typing errors):**
   The database was analyzed looking for data that may have been corrupted, either manually by a typo in the recording of the data, or by merging the tables together. The corrupted data was substituted in case that it was recoverable or dropped. The database was also cleaned up of outliers that may have existed with the objective of avoiding weighted results in specific cases (Section 4.2.2). The missing data was treated differently depending on the type of data contained in the column, in some cases it was filled by a known constant, in extreme cases the row was removed from the database. Both the missing and corrupted data had similar solutions, first we would look the other observations from the same column and see if the missing data was a constant, or a known value, otherwise we would look at related columns, for example, the case with the columns age (*AGE*) and year of birth (*YOBIRTH*) from the VITALS table. If none of the before mentioned solutions worked the row would be removed. The percentage of missing data in the data set can be seen in Figure 3.1 in

which the columns form the database appear from the highest percentage of the missing data to the lowest (left to right), the columns missing from the image did not have unknown data.

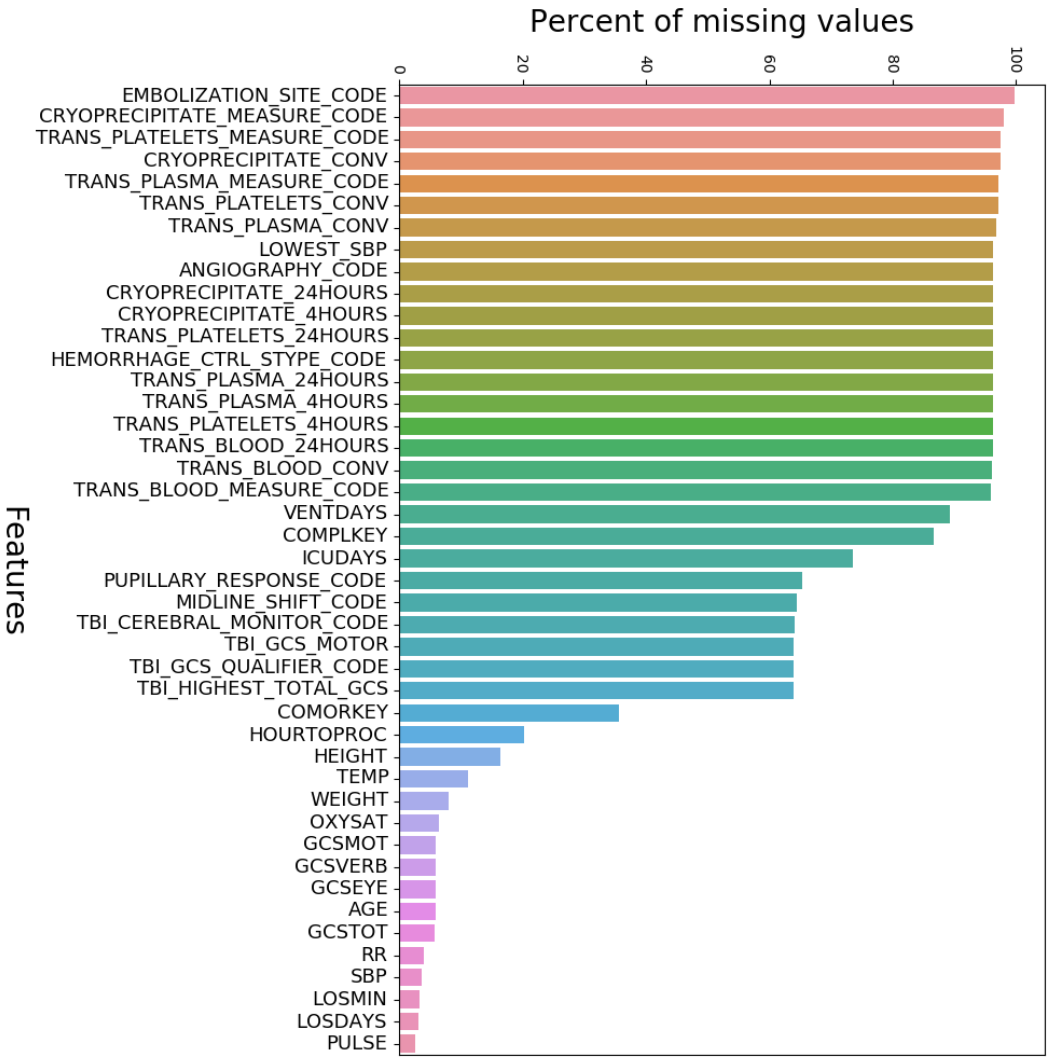


Figure 3.1.: Percentage of highest missing values in the database

4. **Analysis of correlation in the table between variables:**
   A correlation analysis between variables from distinct tables (Section 4.2) has been done in order to find significant features to support the creation of a predictive model. However, after analyzing the results (Section 5.1), we concluded that no significant linear correlation existed between variables.

5. **Evolution graphs of the patients:**
   Proving difficult to find a set of features to create a model, we moved on to generating statistical graphs from the data as an epidemiological study, such as, higher mortality code, higher prevalence of accidents per group age, or gender (Section 4.1).

3.3. Tools and materials

The tools used to carry out this Master Thesis will be detailed below:

Python

Python [65] was chosen as the programming language to act as a foundation of this thesis, supporting the data analysis and processing. Is an interpreted, general-purpose, high-level programming language which was created by Guido van Rossum and whose first version was released in 1991. Its philosophy is mainly based on the readability of the code and supports multiple programming paradigms.

Python is a multi-paradigm programming language. Object-oriented and structured programming are fully supported, and many of its features support functional programming and aspect-oriented programming (including by meta-programming and meta-objects (magic methods)). Many other paradigms are supported via extensions, including design by contract and logic programming.

Scipy

SciPy [66] is a free and open source library for Python. Chosen as a support library thanks to the wide array of statistical functions it contains. It is made up of mathematical tools and algorithms. It was created from Travis Oliphant's original collection, which consisted of Python extension modules, and was released in 1999 under the name Multipack, named for the netlib packages that brought together ODEPACK, QUADPACK, and MINPACK.

SciPy contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE resolution and other tasks for science and engineering.

Pandas

Pandas [67] is a software library written for the Python programming language for data manipulation and analysis. It offers data structures and operations for manipulating numerical tables and time series.

Features of the library:

- DataFrame object for data manipulation with integrated indexing.
- Tools for reading and writing data between in-memory data structures and different file formats.
- Data alignment and integrated handling of missing data.
- Reshaping and pivoting of data sets.

- Label-based slicing, fancy indexing, and subsetting of large data sets.
- Data structure column insertion and deletion.
- Group by engine allowing split-apply-combine operations on data sets.
- Data set merging and joining.
- Hierarchical axis indexing to work with high-dimensional data in a lower-dimensional data structure.
- Time series-functionality: Date range generation and frequency conversion, moving window statistics, moving window linear regressions, date shifting and lagging.
- Provides data filtration.

Pandas was used as a loader of the database into Python. The database was cleaned using Pandas and only the necessary parts to each function, such as demography graphs, were extracted.

SQL

Structured Query Language (SQL) is a domain-specific language used in programming and designed for managing data held in a Relational Database Management System (RDBMS), or for stream processing in a Relational Data Stream Management System (RDSMS). It is particularly useful in handling structured data, i.e. data incorporating relations among entities and variables.

For this Master Thesis, SQLite [68] was used for its simplicity as a standalone database manager, eliminating the need of a client-server connection. It is the most widely deployed database engine (RDBMS), as it is used today by several widespread browsers, operating systems, and embedded systems (such as mobile phones), among others [69].

SQL was used as a tool to merge the tables of the database, as well as query the database when fully loaded, due to Pandas being unable to fully load it, as the database contained a large amount of data.

4. Development

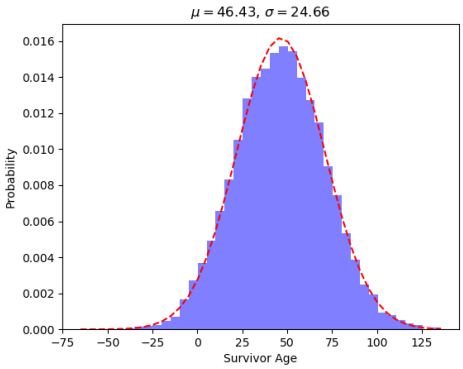
4.1. Data extracted from database

In this Section, the data of the database will be explored looking for parameters or indicators that will provide information to support trauma treatment.

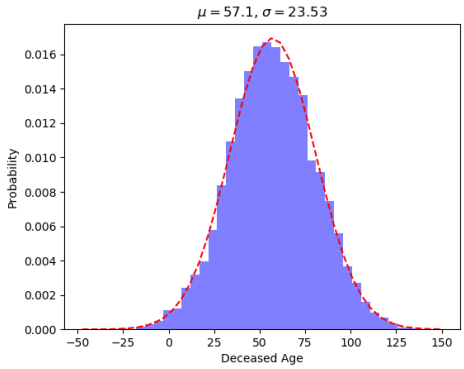
4.1.1. Global data

First of all, as we can see in Figure 4.1, we can observe an age distribution towards middle-age people, with a global mean of 46,69 years old. When looking at survivor and deceased data, we can observe a leaning towards younger people in the survivor group, whereas the deceased group tends towards elder people.

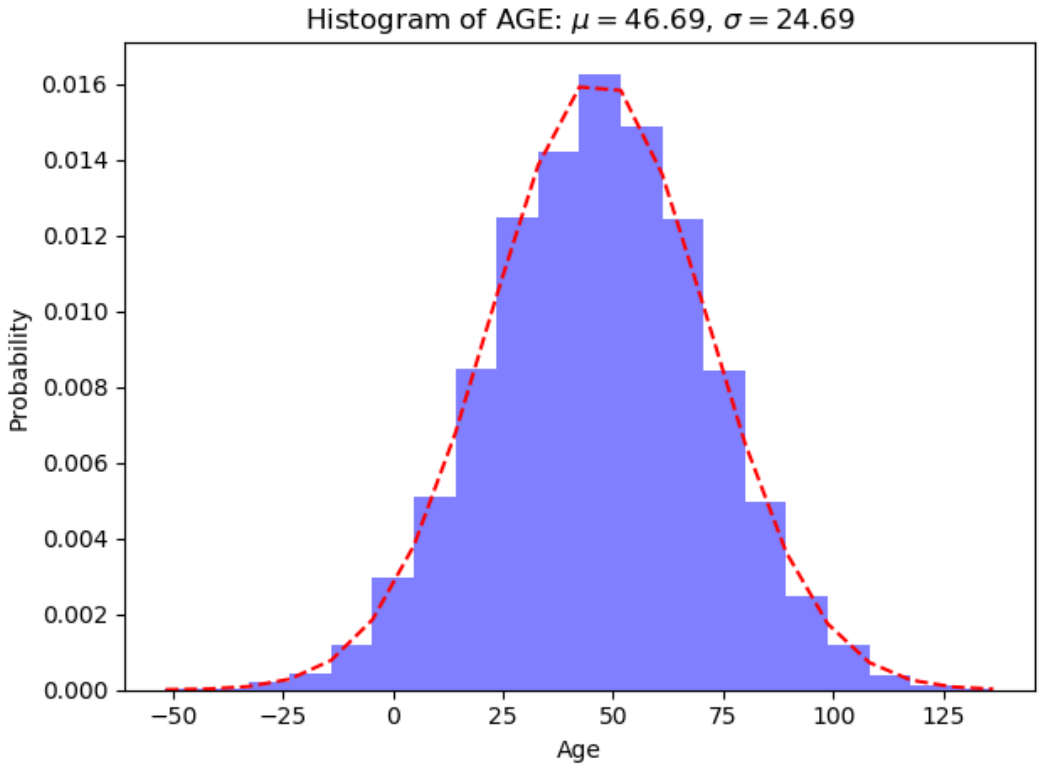
Figure 4.2 shows an histogram of the time taken for the EMS to get to the scene, and to take the patient to a hospital. In both figures the triage times (Section 2.1.7) can be easily seen, especially when observing Figure 4.2b where there is a peak at the lowest values representing the most critical patients. Fewer breaks can also be seen, this is caused as critical patients have a higher priority. In Figure 4.2a more breaks are present, and the peak is in approximately 50 minutes. These two histograms show the importance of quick response time at accidents, and that triage techniques works and are effectively used.



(a) Distribution of age of survivors in the database

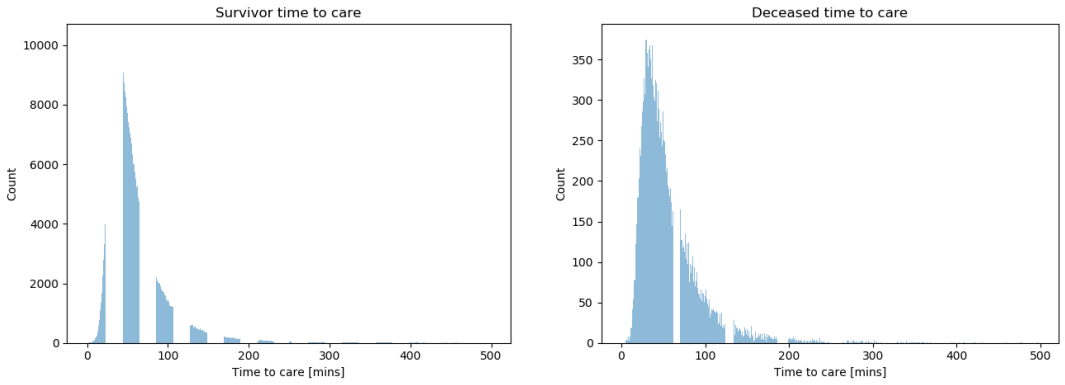


(b) Distribution of age of deceased in the database



(c) Distribution of ages from all patients in the database

Figure 4.1.: Distribution of age in the database

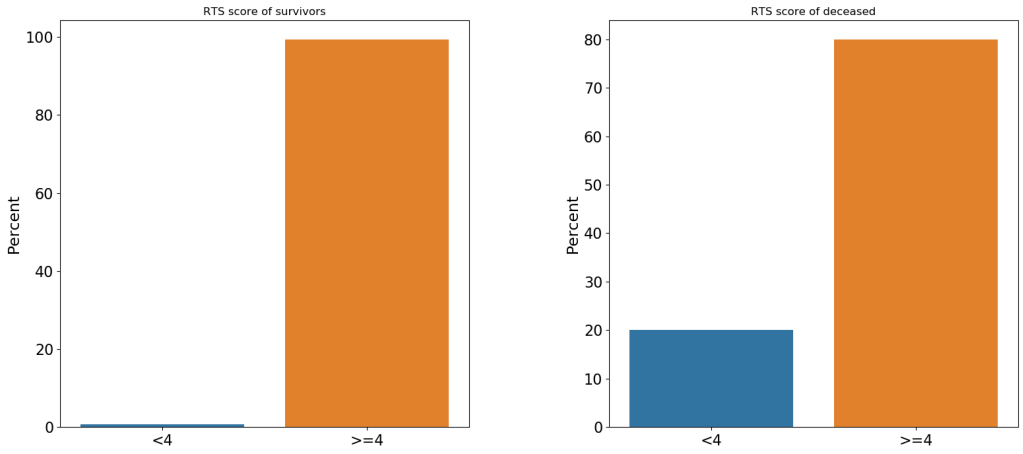


(a) Survivors **(b)** Deceased

Figure 4.2.: Time taken for the EMS to transport the patient to the ED

4.1.2. Trauma scores

In Figure 4.3 we can see that the RTS values equal or higher than 4 represent a positive outcome as seen in Figure 4.3a where the survivor probability is at approximately 90%, while a negative outcome (value lower than 4) is at around 3% of survival (a score of less than 4 drops the survival rate to 50%). However, in Figure 4.3b, whereas the deceased probability is higher than 20% for patients with a RTS score less than 4, the probability of death is at 80% for values considered safe. This will be further studied in Subsection 4.1.3.

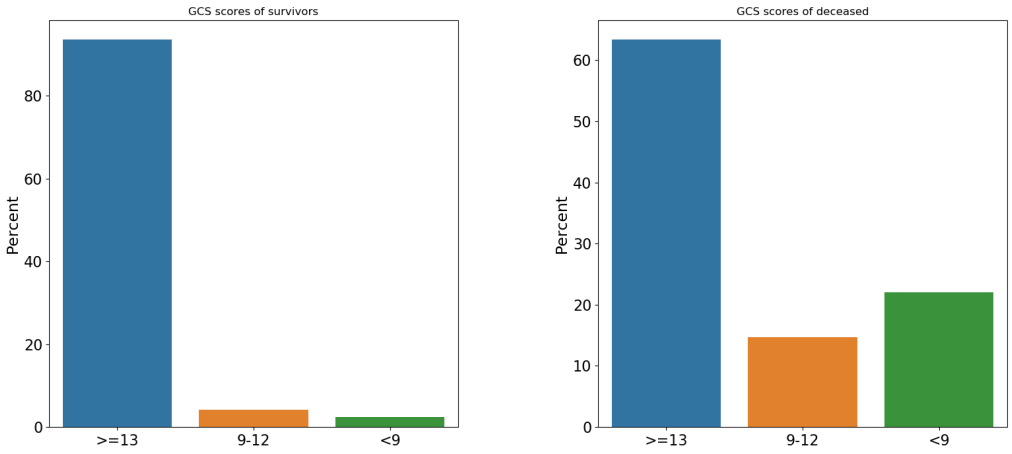


(a) Survivors RTS **(b)** Deceased RTS

Figure 4.3.: Percentage of RTS scores in all patients of the database

In Figure 4.4 we can see that the GCS works better as a determinant of a negative outcome than as a determinant of a positive outcome of patients. In Figure 4.4a the higher the GCS value is the better. A value higher or equal than 13 is considered a minor trauma injury, a value between 9 and 12, moderate, and lower than 9, severe.

However, when looking at Figure 4.4b this no longer applies directly. Although for GCS values from 9 to 12, and lower than 9 groups increased in mortality, especially lower than 9; for a GCS value higher than 13, a value considered safe, it also has a great mortality at more than 60%.

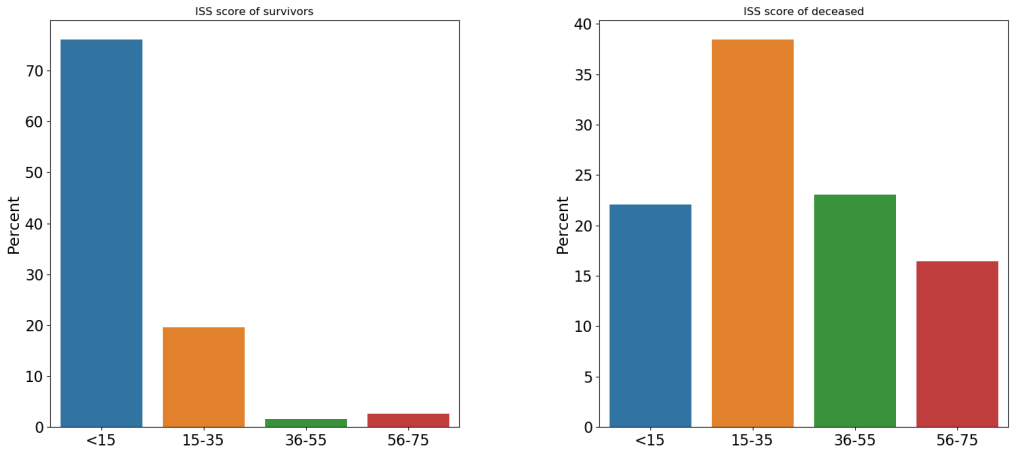


(a) Survivors GCS **(b)** Deceased GCS

Figure 4.4.: Percentage of GCS scores in all patients of the database

The ISS scores in Figure 4.5 are more balanced, however the same problem that was detected for RTS and GCS scores is present in the ISS analysis. In Figure 4.5a the higher probability of survival is under the group of less than 15 ISS score (a score greater than 15 is considered a major trauma), while 15 to 35 has also a notable probability (20%), the other two groups probabilities are much lower. Although is worth noting that for ISS between 56 and 75, which is considered highly critical, has a probability of almost 5% of survival.

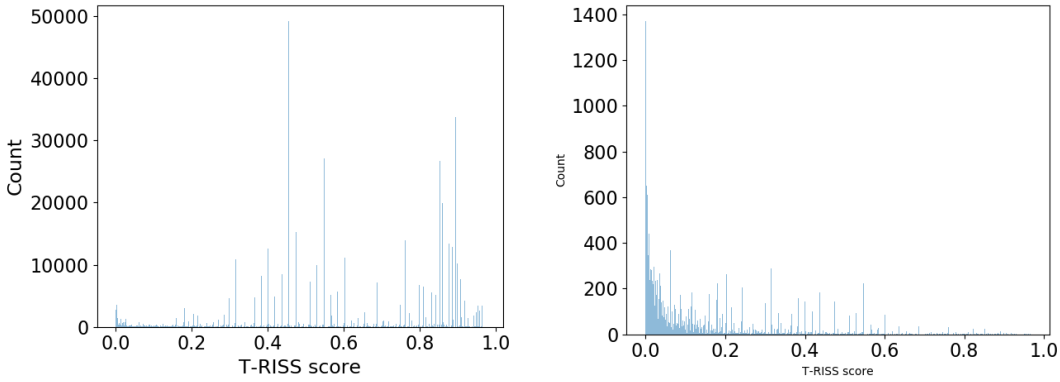
On the contrary, in Figure 4.5b the data is more balanced, with a peak at 15 to 35 indicating critical patients with negative outcomes, and like with RTS and GCS, the stable value has also a high probability of a negative outcome for the patients.



(a) Survivors ISS **(b)** Deceased ISS

Figure 4.5.: Percentage of ISS scores in all patients of the database

Finally, in Figure 4.6 we have the histograms for the TRISS score. This score is the best one performing out of the scores tested and analyzed. Although there are peaks in the middle of the histograms (45-50%), both Figures 4.6a and 4.6b show data accumulated in their respective expected values. Those are, for Figure 4.6a the data is more concentrated in a survival probability higher than 80%, while in Figure 4.6b the concentration occurs below 20% survival probability.



(a) Survivors TRISS **(b)** Deceased TRISS

Figure 4.6.: Count of TRISS scores in all patients of the database

4.1.3. Anomalies in trauma scores

As it could be seen in the previous Subsection, the scores analyzed had a high mortality despite their values being highly positive, normally indicating a good patient outcome. In this Subsection we are going to analyze this problem and to explain its causes.

The scores were calculated from the VITALS table, which included the parameters Systolic Blood Pressure (SBP), Respiratory Rate (RR), pulse, temperature and oxygen saturation among other. They were used to calculate the RTS, ISS, and TRISS scores. The GCS was already calculated in the VITALS table. Giving a thorough look at the data we noticed that, per patient, we only had two records of this data, one issued by the EMS and other by the ED.

The first problem that was noticed is that by having only these two records we only had the information of the patient when he or she is attended by the emergency services. Furthermore, this meant that if a patient were to deteriorate after being admitted in the ED, we would not have data about it, and therefore, the scores were unable to show an evolution of the patient.

Another problem found was that while the ED had all their data correctly, the EMS data was missing for a lot of patients, this was a problem in two ways. The first problem was that it affected data density as there were only two records, and this meant losing one for a lot of patients. The other one is that either the data was corrupted, or the EMS were not logging correctly their data. This problem can be seen in Table 4.1.

Total incomplete records	28,82%
EMS incomplete records	24,25%
ED incomplete records	4,57%

Table 4.1.: Incomplete or corrupt records from VITALS

For the first problem we used the COMPLIC (it contains data for the complications suffered by the patients in the ED) table to try and determine whether complications inside the ED were the cause of the anomalies. In Figure 4.7 we can observe that in more than 50% of the patients with a RTS score higher than 7 had a complication in their stay.

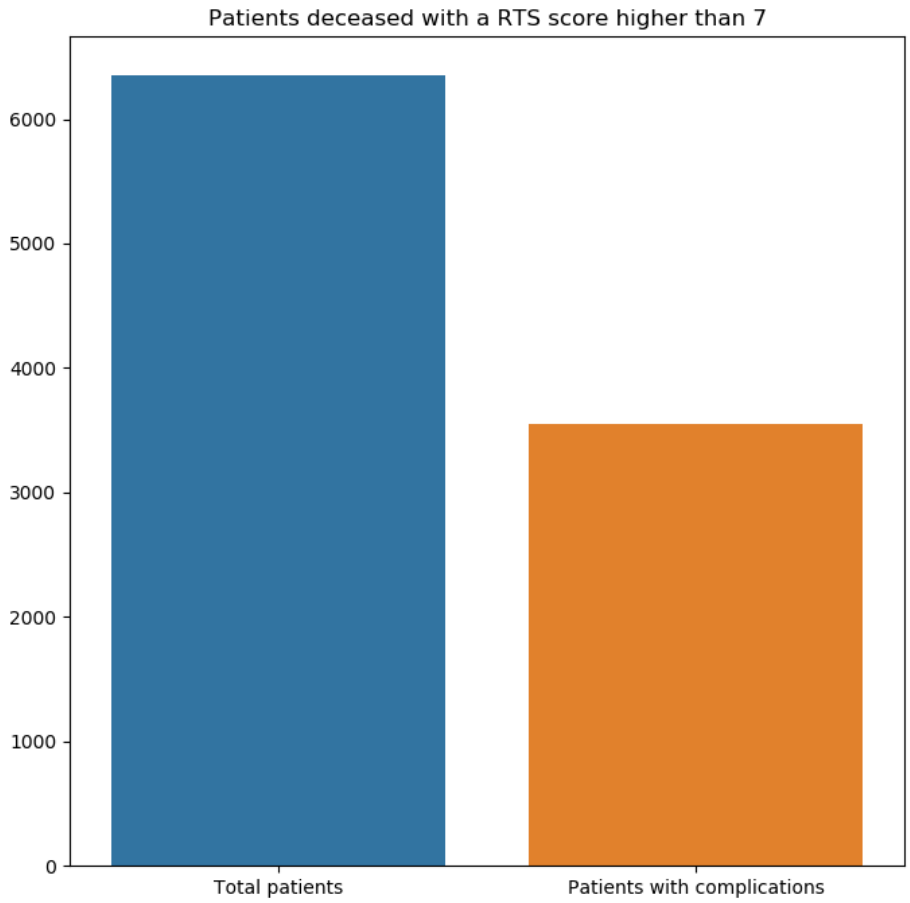


Figure 4.7.: Comparison of RTS scores depending on conditions

Moreover, we calculated the specific complications that patients with a RTS of more than 7 suffered. The results can be seen in Figure 4.8. Cardiac arrest and unplanned intubation are the two most common complications, followed by an unplanned entry to the ICU, and acute kidney injury. The top three complications make sense as a RTS of 7 or higher is unlikely to have problems.

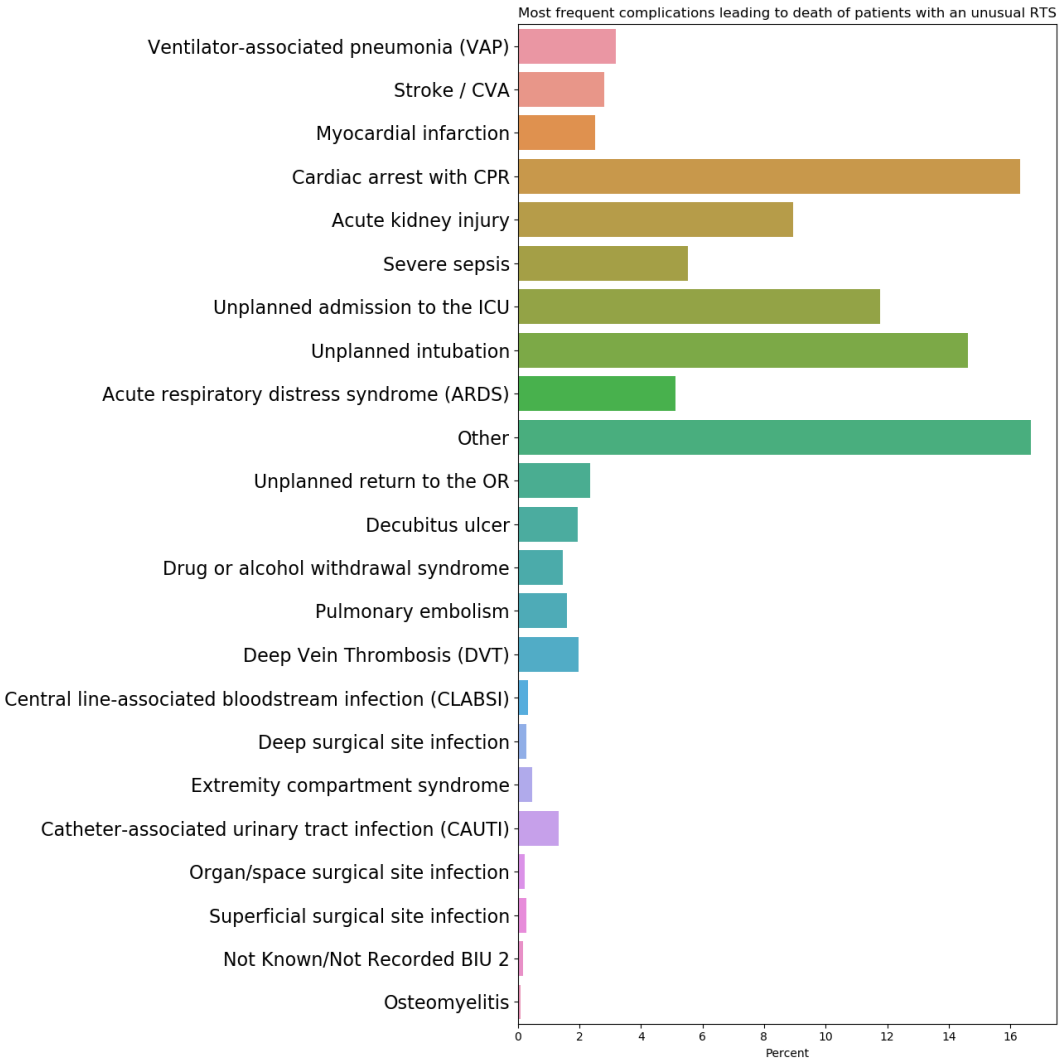


Figure 4.8.: Complications in patients with RTS score higher than 7

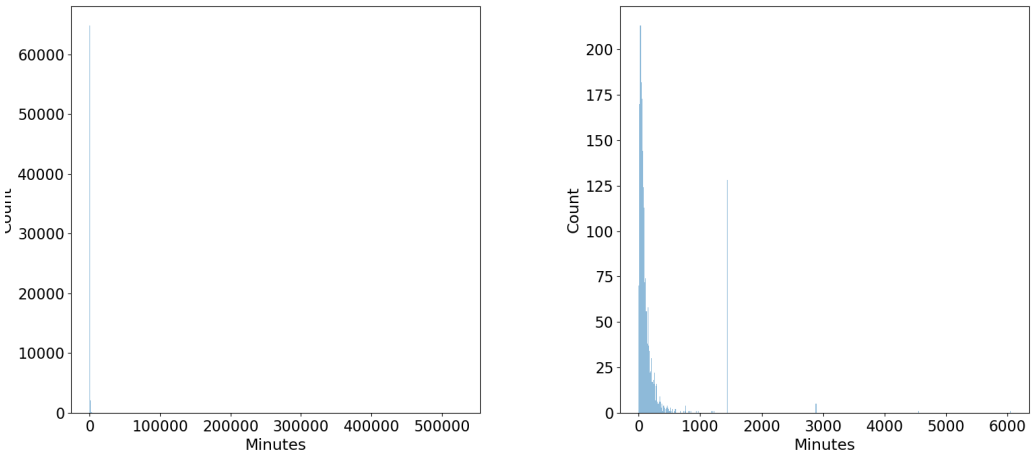
To conclude this Subsection, as seen, the most probable problem that caused positive values in trauma scores to have high rate of death are complications during care. Moreover, as we only have one data entry, two in some cases, we cannot define an evolution of the trauma scores in patients.

4.1.4. Patient vital signs with trauma

The objective of this Subsection is to analyze patients' vital signs in order to learn the importance of these parameters in the outcome of an injured patient.

Firstly, in Figure 4.9 we have analyzed the time spent in the ED by the patients. One quick look over both histograms makes clear that survivors spend almost no time in the ED (Figure 4.9a). Nevertheless, for deceased patients the results are much higher, with patients staying up to 8 hours with a peak at 24 hours (Figure 4.9b). The reason for this peak is that while other records were completed using a

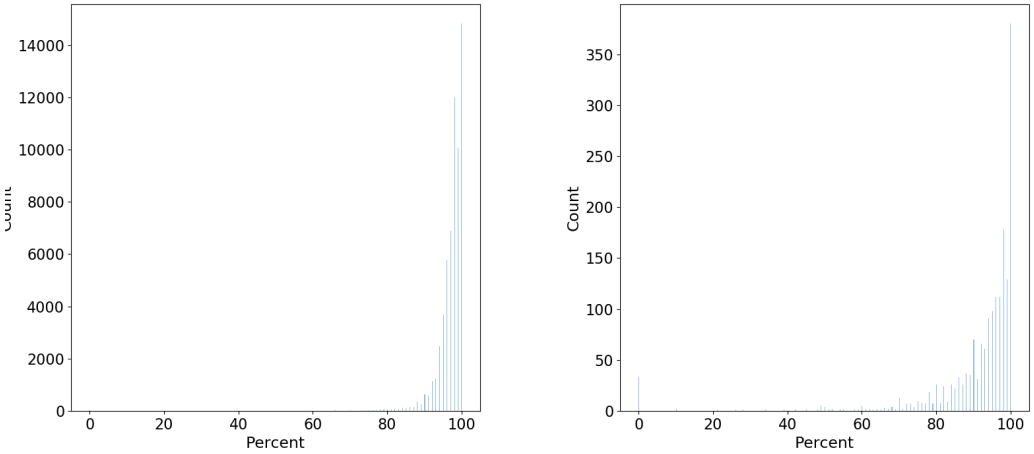
field in table ED called *EDMINS*, these records were computed from another field called *EDDAYS*, this caused that if *EDDAYS* were 1 day, *EDMINS* would become 1440 minutes, as there is no data in *EDMINS* in those cases.



(a) ED time of survivors **(b)** ED time of deceased

Figure 4.9.: ED times of patients in the database

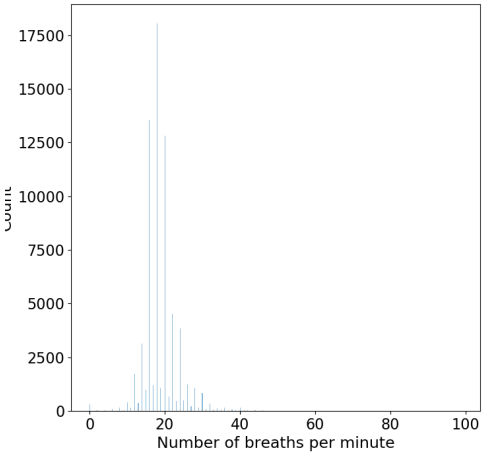
Figure 4.10 shows the difference in oxygen saturation between survivors and deceased patients. We can observe how in the case of survivors (Figure 4.10a) the oxygen saturation in few cases drops below 90%. However, in the case of deceased patients, although higher saturation levels are existent (same problem with the trauma scores in Subsection 4.1.3), the data is more dispersed with a higher number of values below 90% (Figure 4.10b). Values of oxygen saturation below 90% are considered critical hypoxemia and/or hypoxia [70, 71].



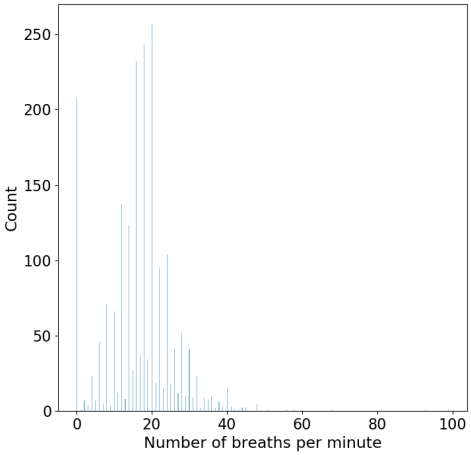
(a) Oxygen saturation of survivors **(b)** Oxygen saturation of deceased

Figure 4.10.: Oxygen saturation of patients in the database

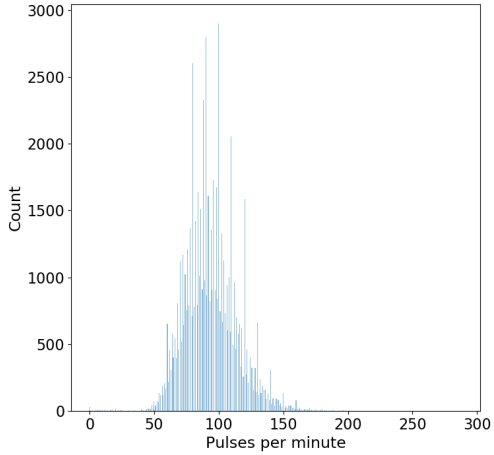
We have joined the Respiratory Rate (RR) (Figures 4.11a and 4.11b), the pulse (Figures 4.11c and 4.11d), and the Systolic Blood Pressure (SBP) (Figures 4.11e and 4.11f). As it can be seen in non-critical patients these values stay withing normal ranges, a RR of approximately 20 breaths/min, a pulse of approximately 200 beats per minute, and a SBP of around 120 mm Hg. However, in the critical cases these values get more distributed, rather than concentrated, in their respective histograms. As the person goes into shock the RR increases, with values over 40, as does the pulse and the SBP. On the contrary, as the person goes into cardiac arrest these values decrease dramatically, with peaks at 0.



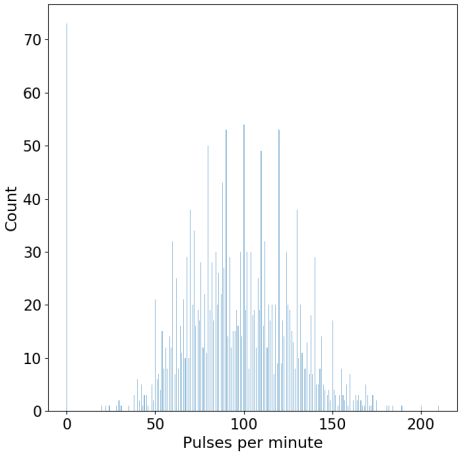
(a) Respiratory rate of survivors



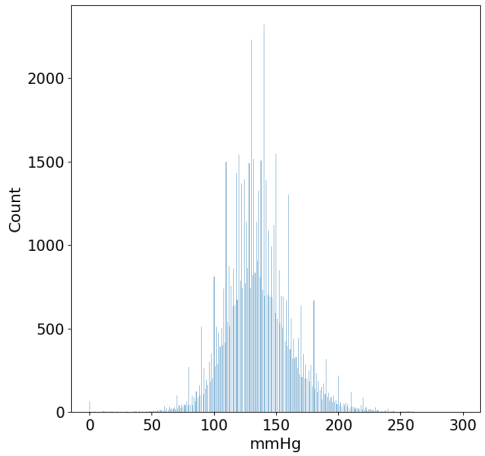
(b) Respiratory rate of deceased



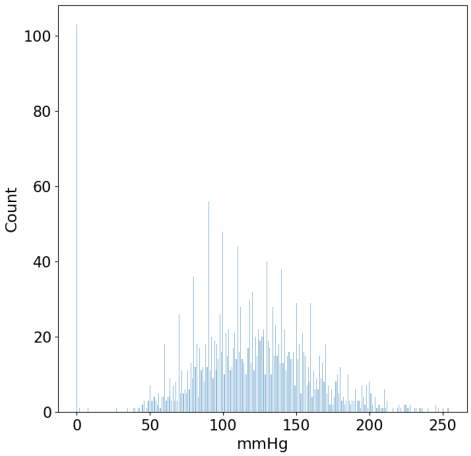
(c) Pulse of survivors



(d) Pulse of deceased



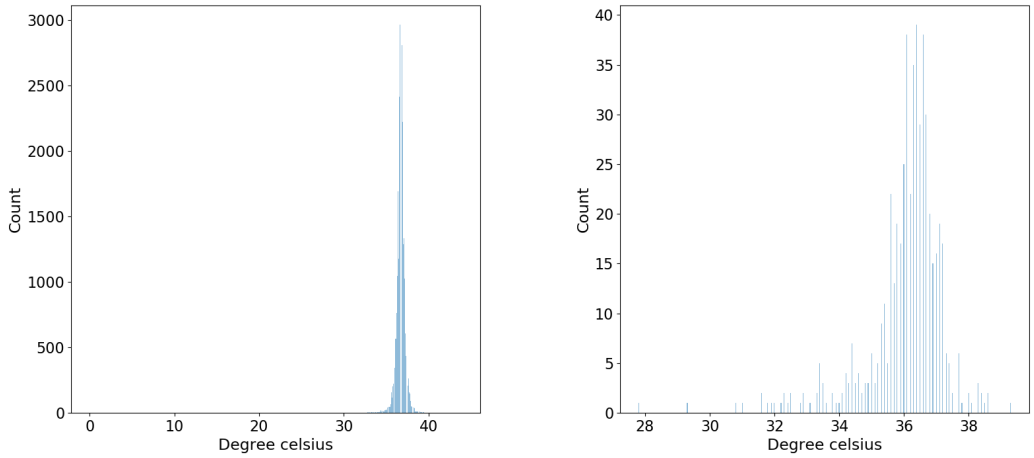
(e) SBP of survivors



(f) SBP of deceased

Figure 4.11.: Respiratory rate, SBP, and pulse of patients in the database

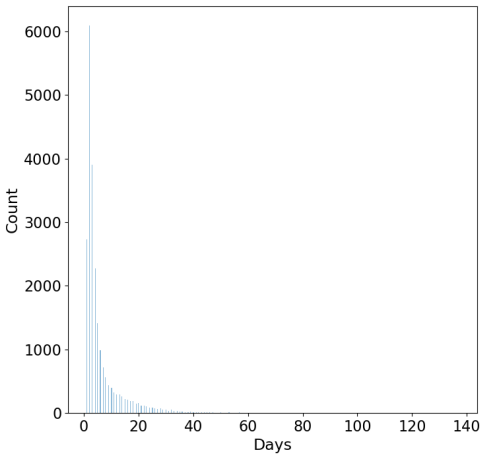
In Figure 4.12 we can observe how the status of the patient affects the body temperature. While in stable cases the temperature sits at around 36 degrees Celsius with a defined normal distribution concentrated in that value, in critical cases the distribution is more spread, going towards hypothermia. One probable cause for this is by manually induced hypothermia or Targeted Temperature Management (TTM), in [72], it shows that in situations with low blood pressure reducing the temperature of the patient can prevent lasting damage to critical organs, such as the brain. As the temperature drops, the body functions become slower, therefore using less blood.



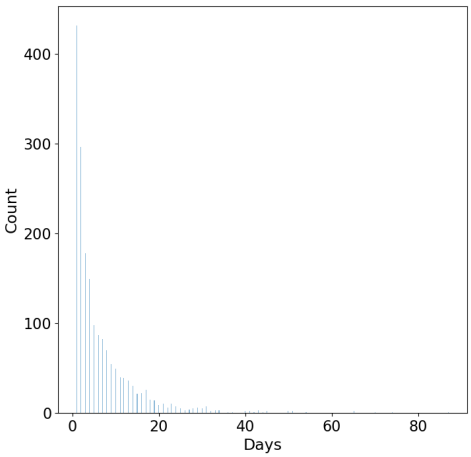
(a) Temperature of survivors **(b)** Temperature of deceased

Figure 4.12.: Temperature of patients in the database

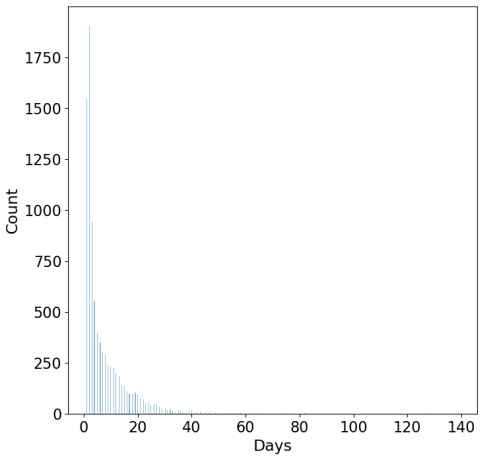
Finally, we have the ICU and ventilator times in hospital in Figure 4.13. Both the histograms show a very similar curve. This could be explained as some studies [73] have shown that a mean of 39,5% (± 15,2) of ICU patients are mechanically ventilated. This is further confirmed in Subsection 5.1 where a correlation between days in ICU and days with mechanical ventilation is found.



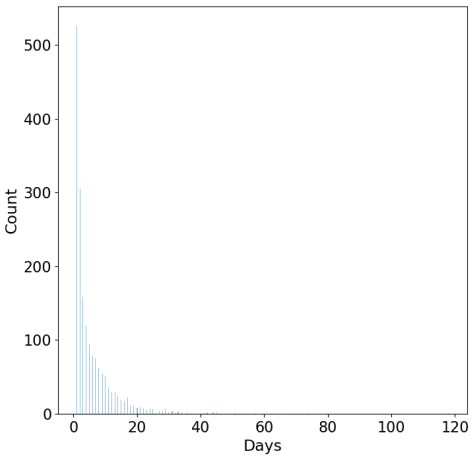
(a) ICU time of survivors



(b) ICU time of deceased



(c) Ventilator time of survivors



(d) Ventilator time of deceased

Figure 4.13.: ICU and ventilator time of patients in the database

However, as we could see, the same problem as the one faced with trauma scores had with the missing data from the EMS is also present in this analysis, as the data is extracted from the same table (VITALS).

4.1.5. Car accidents

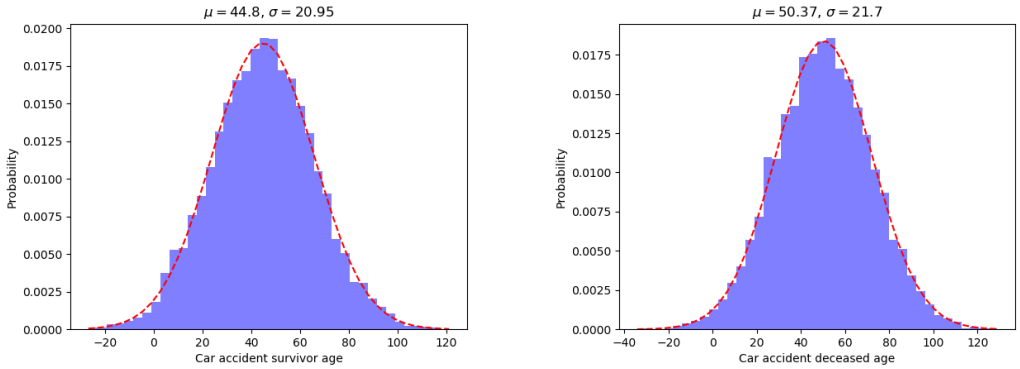
One of the most frequent and mortal traumas in the database are car accidents, in the category injuries to driver, the companion, or pedestrians are included. In Table 4.2 we can observe the mean of ages related to car accidents.

As we can see in Table 4.2, older people is more prone to critical outcomes from this type of traumas. In Figure 4.14 we can see the distribution of this relation and peaks at 50 to 60 years old can be seen. Furthermore, although smaller, there is a peak of deceased between 20 and 30 years old, indicating that young adults may take

	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
General	44,8	20,95	50,37	21,07
Male	47,37	20,8	46,04	22,04
Female	44,36	20,79	44,37	21,89

Table 4.2.: Age statistics in car accidents

more risks when driving.



(a) Distribution of age of survivors of car accidents

(b) Distribution of age of deceased of car accidents

Figure 4.14.: Distribution of age in survivors and deceased in car accidents

4.1.5.1. Trauma injuries from car accidents

We have analyzed the most common traumas occurred in car accidents and chosen the 5 most common to further study. The most common traumas in car accidents are represented in Figure 4.15.

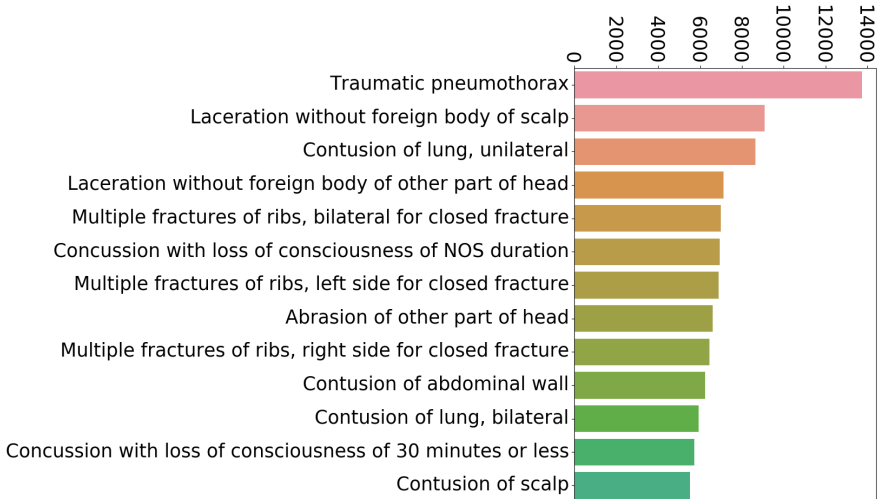


Figure 4.15.: Most common traumas in car accidents

The 5 selected to be further analyzed were:

- Traumatic pneumothorax.
- Multiple fractures.
- Laceration without foreign body of scalp.
- Lung contusion.
- Traumatic subarachnoid hemorrhage.

Their respective mortality count in car accidents scenarios can be seen in Figure 4.16.

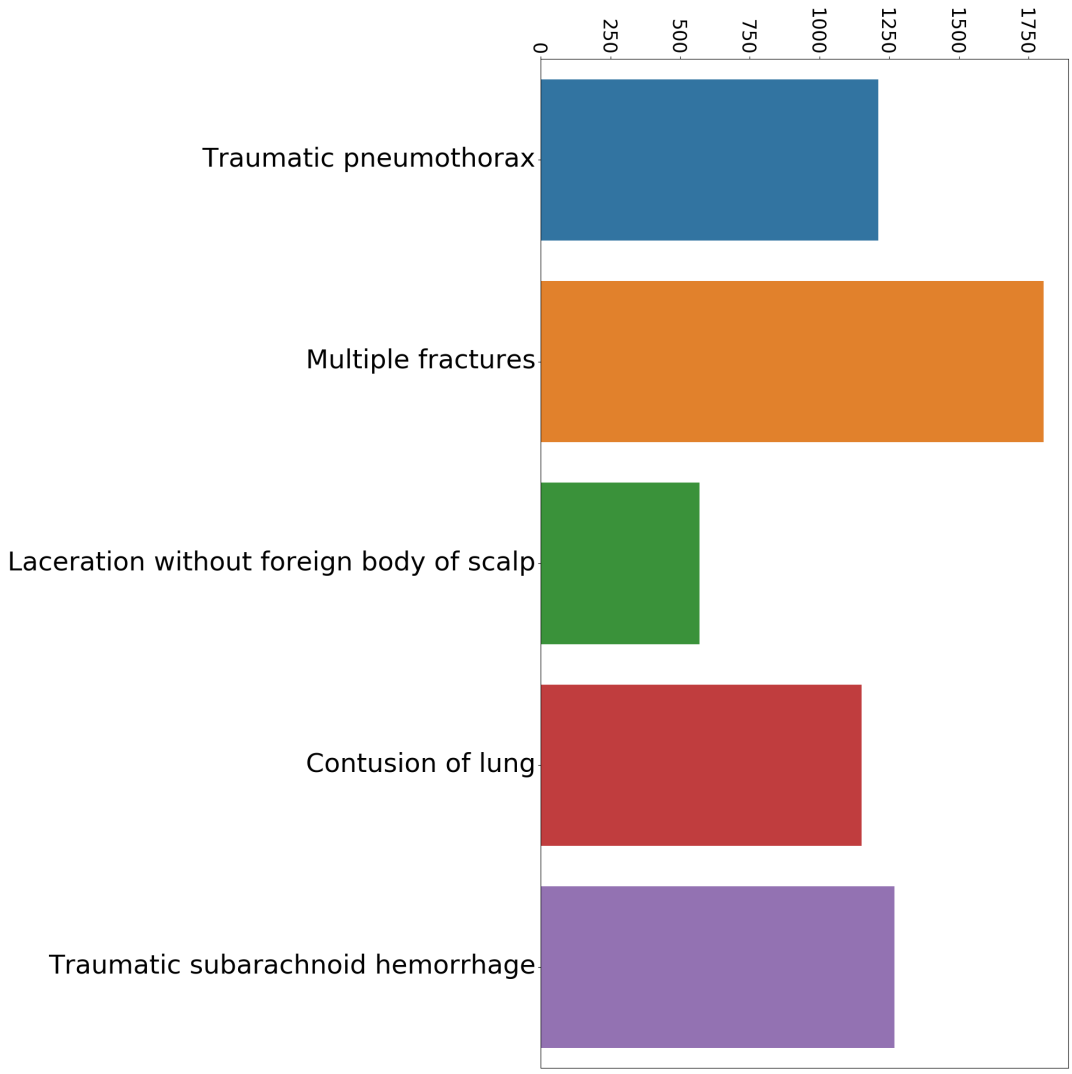


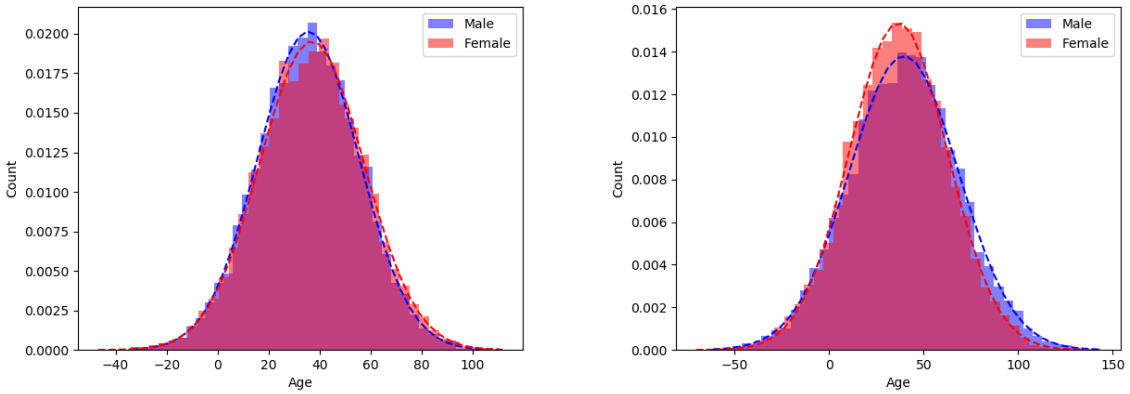
Figure 4.16.: Mortality of traumas in car accidents

In Table 4.3 the mean age for the injuries mentioned in Table 4.16 can be seen. In lung contusion, although their mean values are similar when looking at the distribution (Figure 4.17), differences can easily be seen, the peak of age is younger in the case of men compared to the peak of age in female patients. The other traumas present a similar distribution.

Traumatic pneumothorax				
	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
Male	49,83	20,79	50,67	22,04
Female	44,33	20,78	49,83	21,89
Lung contusion				
	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
Male	44,33	20,79	50,49	22,09
Female	44,33	20,78	49,32	21,99
Laceration				
	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
Male	44,35	20,79	49,37	22,41
Female	44,35	20,79	45,47	21,11
Multiple fractures				
	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
Male	44,34	20,79	51,27	21,95
Female	44,35	20,79	50,7	21,36
Traumatic subarachnoid hemorrhage				
	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
Male	44,34	20,79	51,43	22,1
Female	44,34	20,79	50,75	21,09

Table 4.3.: Age statistics in traumas from car accidents

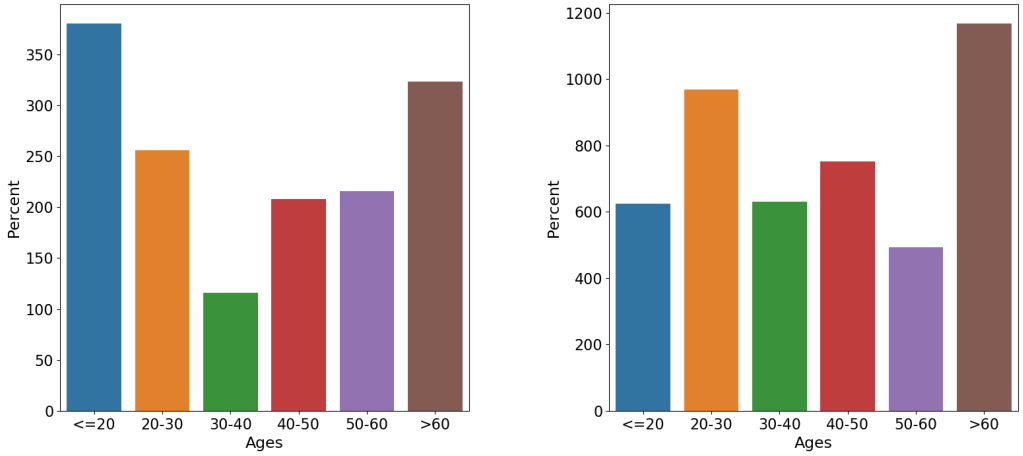
As shown in Figure 4.17 a certain difference can be seen between the two groups of population with lung contusion. In the case of survivors both distributions are almost equal, with the same exact mean and a very close deviation. However, in the deceased patients population, women have a higher probability to die at younger ages than men, whereas male patients have a higher mortality in elderly patients.



(a) Survivors **(b)** Deceased

Figure 4.17.: Comparison of the distribution of age in patients with lung contusion

In Figure 4.18 we can observe how the difference in distribution of the previous histograms is caused by a difference in the count of patients by gender. In Figure 4.18a the population count peaks in teens and old adults, while in Figure 4.18b the population tends more towards middle-aged adults, although after elder people the most predominant is young adults.



(a) Groups of age with lung contusion in deceased female patients **(b)** Groups of age with lung contusion in deceased male patients

Figure 4.18.: Comparison of group ages count between genders

To conclude this Subsection, both patient vital signs and trauma scores were equal to the ones exposed in Subsection 4.1.4 and Subsection 4.1.2 respectively.

4.1.6. Falls on same level

Another trauma we wanted to focus on are falls on same level as it represents the second most common cause of death. This represents the falls without falling more than the patient's height. In Table 4.4 we can see the mean of age in these situations. As it can be easily seen, elderly people are more prone to negative outcomes.

	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
General	44,75	20,91	75,93	10,6
Male	42,67	19,52	74,62	11,24
Female	49,32	23,01	77,8	9,3

Table 4.4.: Age statistics in falls on same level

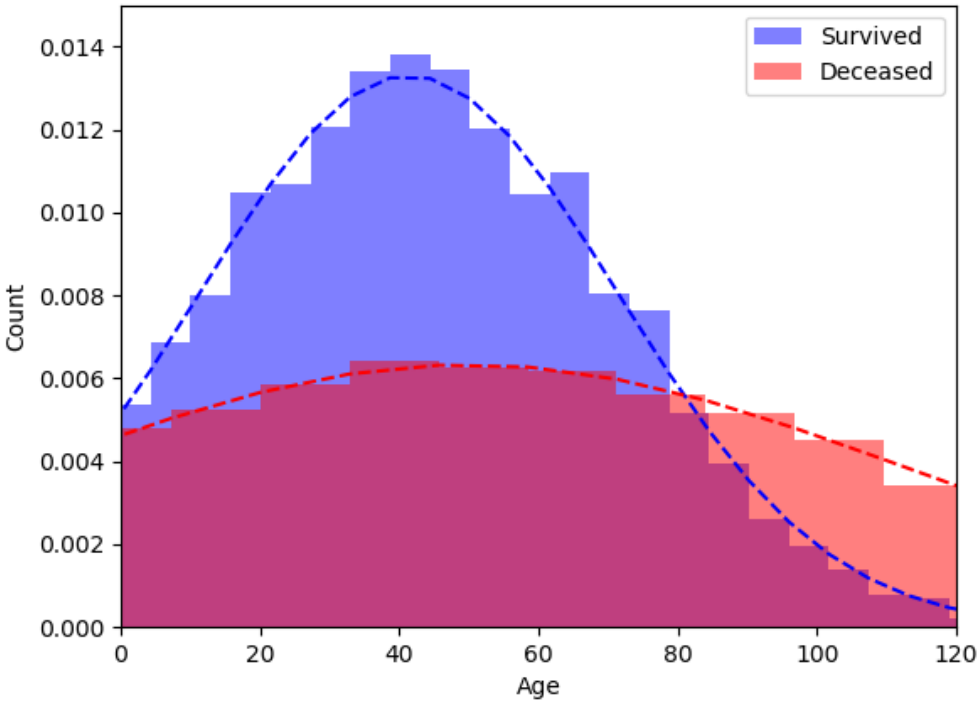


Figure 4.19.: Comparison of the distribution of age in survivors and deceased in falls

That phenomena can be graphically seen in Figure 4.19 where the peak for survivors is around 40, and the peak for deceased is around 60 years old. The distribution is also trending to older patients. However, two other peaks exist in the survived histogram,

one is young adult, this population is more prone to falls in sports or accidents, than in just walking [74]. Furthermore, the second peak is in elderly patients with more than 60 years old, this population is more prone to falls caused by weak bone health [75].

4.1.6.1. Trauma injuries from falls on the same level

For this trauma we have also analyzed the most common injuries that are produced by falls on same level, and chosen 5 of them to study more closely. The most common traumas caused by falls on same level can be seen in Figure 4.20.

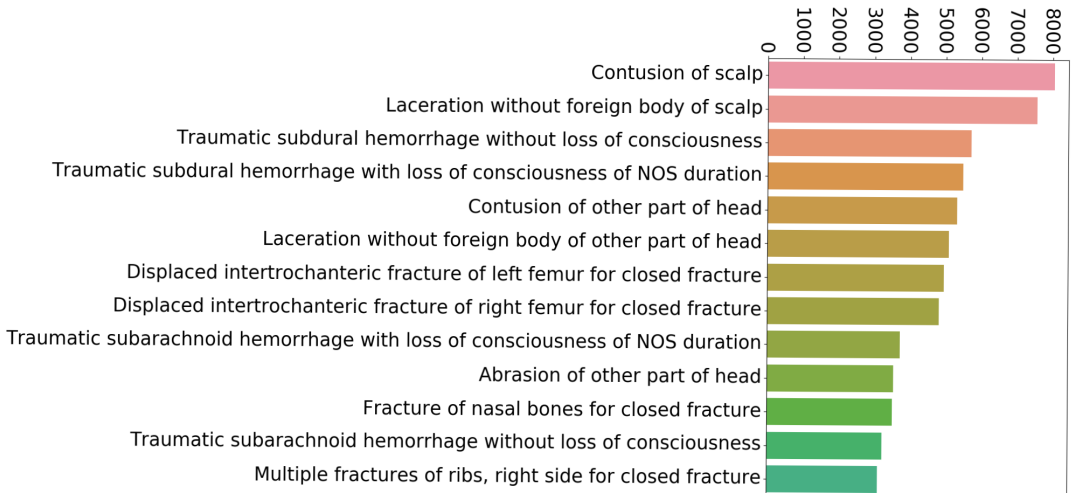


Figure 4.20.: Most common traumas in falls in same level

The traumas chosen to further study were the following:

- Contusion of scalp
- Laceration without foreign body of scalp
- Traumatic subdural hemorrhage
- Traumatic subarachnoid hemorrhage
- Traumatic hemorrhage of cerebrum

The 5 of them are injuries that occur in the head, this is because the head is the most damaged part in the body when a fall on same level takes place (62.2% of the total of injuries sustained in this trauma) [76]. The mortality count of these traumas can be observed in Figure 4.21.

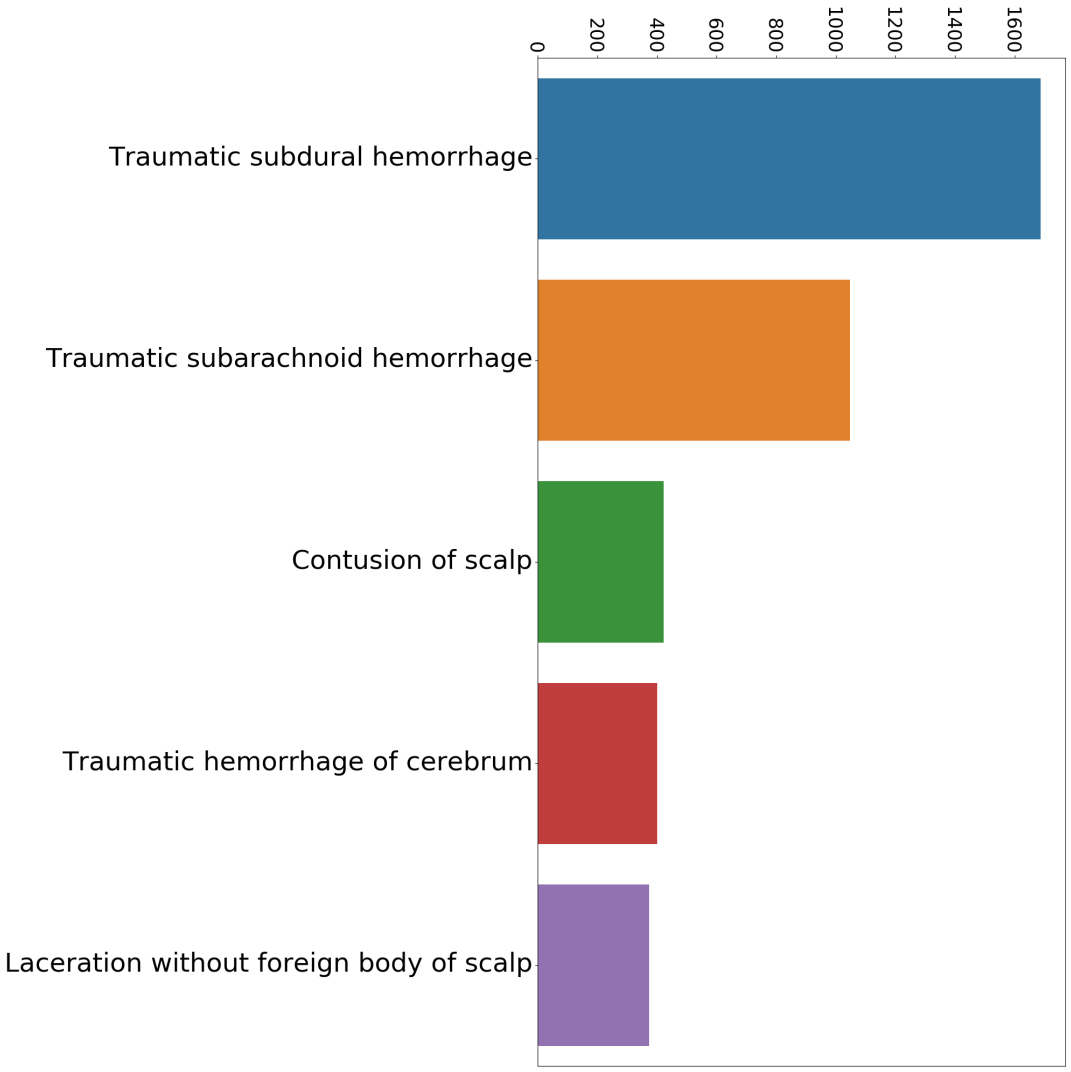


Figure 4.21.: Highest mortality traumas in falls in same level

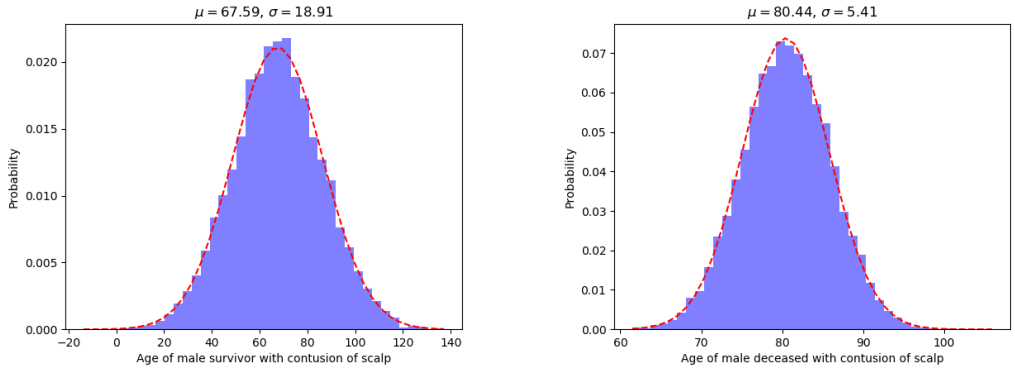
In Table 4.5 we can see the age distribution of the 5 injuries. At a first glance, it is easy to determine that the population trends towards old people. As the traumas studied causes damage to the head, the population with a higher prevalence to this type of injuries are older, as opposed to the case with younger populations which have a higher risk of sustaining bone fractures[76].

Moreover, although in most of the traumas the mean age for survivors and deceased are close together, the standard deviation in deceased patients is really low compared to the case of positive outcomes. This means that while both means are together, survivors are more dispersed in the age spectrum, but the deceased patients are more concentrated towards older ages. Nevertheless, in the case of scalp contusion the age difference increases having a survived mean of 67,59 $\pm$ 18,91 years old, and a deceased mean of 80,44 $\pm$ 5,41 years old.

Traumatic subdural hemorrhage				
	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
Male	68,22	16,28	73,92	11,59
Female	74,07	12,78	76,92	8,76
Scalp contusion				
	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
Male	67,59	18,91	80,44	5,41
Female	73,8	14,78	76,82	9,8
Traumatic hemorrhage of cerebrum				
	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
Male	71,96	15,76	70,4	10,72
Female	67,67	13,46	73,48	11,93
Laceration				
	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
Male	66,35	19,1	74,97	10,86
Female	74,68	13,5	75,97	7,8
Traumatic subarachnoid hemorrhage				
	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
Male	67,96	15,29	70,3	12,41
Female	73,37	13,5	77,03	9,72

Table 4.5.: Age statistics in traumas from falls on same level

In Figure 4.22 we can better see the difference between survivors of scalp contusion (as this injury presents a higher age difference than the rest) in falls on same level (Figure 4.22a), and deceased (Figure 4.22b). In the first group the peak is above approximately 75 years old, as for the deceased patients the peak is over 80 years old.

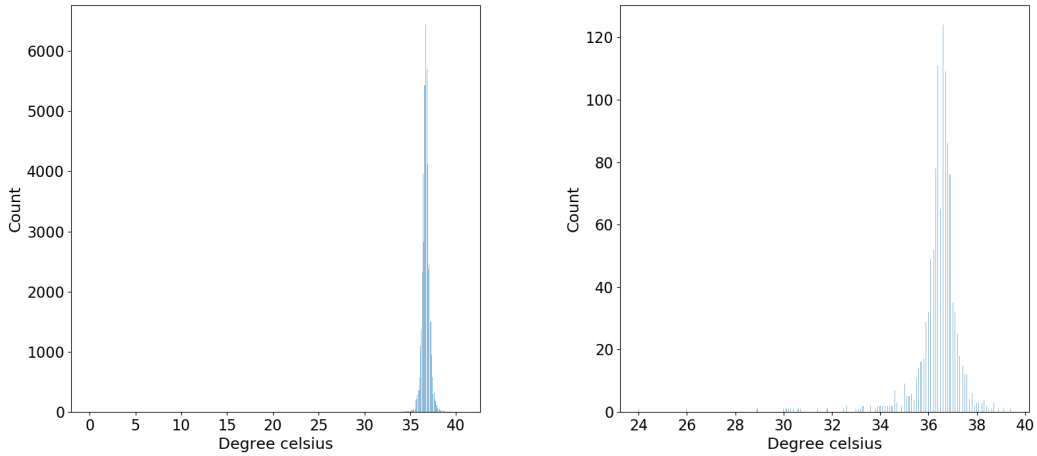


(a) Survived **(b)** Deceased

Figure 4.22.: Distribution of age of survivors in falls on same level with scalp contusion

4.1.6.2. Difference of patient vital signs in falls on same level

Although all vital parameters are very similar to that of the general population of the database (Subsection 4.1.4), the temperature in the case of this trauma (Figure 4.23) differs from the temperature analyzed in the previous Section (Subsection 4.1.4) in Figure 4.12. In this case there is not a general hypothermia, this may be caused by the fact that while treating elderly patients that practice may be dangerous.



(a) Temperature of survivors **(b)** Temperature of deceased

Figure 4.23.: Temperature of patients in falls on same level

4.1.7. Motorcycle accidents

The last of the 3 traumas that we are analyzing are motorcycle accidents. As with the car accidents this includes driver, companion, and pedestrians. In Table 4.6 the

age distribution can be observed.

	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
General	44,8	20,95	50,37	21,07
Male	47,37	20,8	46,04	22,04
Female	44,36	20,79	44,37	21,89

Table 4.6.: Age statistics in car accidents

Looking at the age statistics of motorcycle accidents, it can be seen that the distribution in the general population, as well as in both genders, are very similar. Moreover, both survivors means and deceased means almost have no difference between each other.

Furthermore, when extracting data for this type of trauma, we had some problems with traumas not having female patients. In Figure 4.24 the distribution of gender in motorcycle accidents can be seen. Female patients represent approximately 11% of the total population of this trauma.

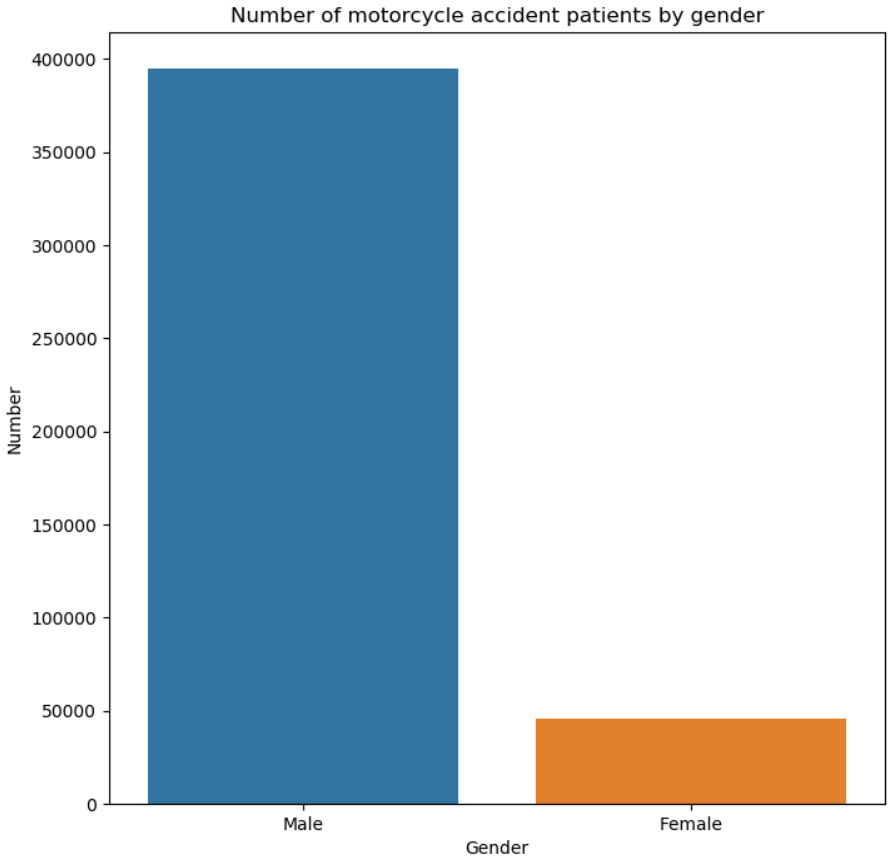


Figure 4.24.: Number of motorcycle accidents by gender

4.1.7.1. Trauma injuries from motorcycle accidents

As with the two previous traumas, we have also analyzed the most common injuries that patients show in motorcycle accidents, and chosen 5 of them. The most common injuries caused by motorcycle accidents can be seen in Figure 4.25.

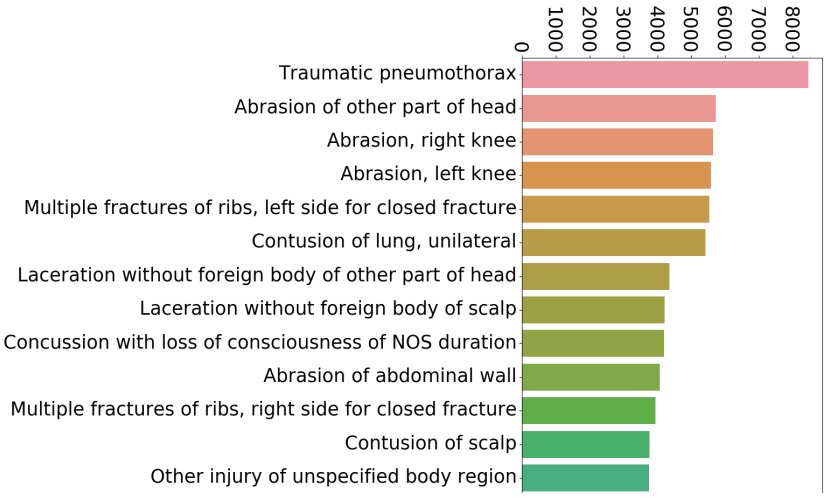


Figure 4.25.: Most common traumas in motorcycle accidents

We did not include the traumatic pneumothorax injury due to it being the same as the case with the car accidents. The traumas chosen to study were the following:

- Abrasion
- Concussion
- Laceration
- Multiple fractures
- Lung contusion

Abrasion and laceration are the two most mortal of the chosen traumas, with abrasion having caused more than 70.000 deaths in motorcycle accidents. Their mortality can be seen in Figure 4.26.

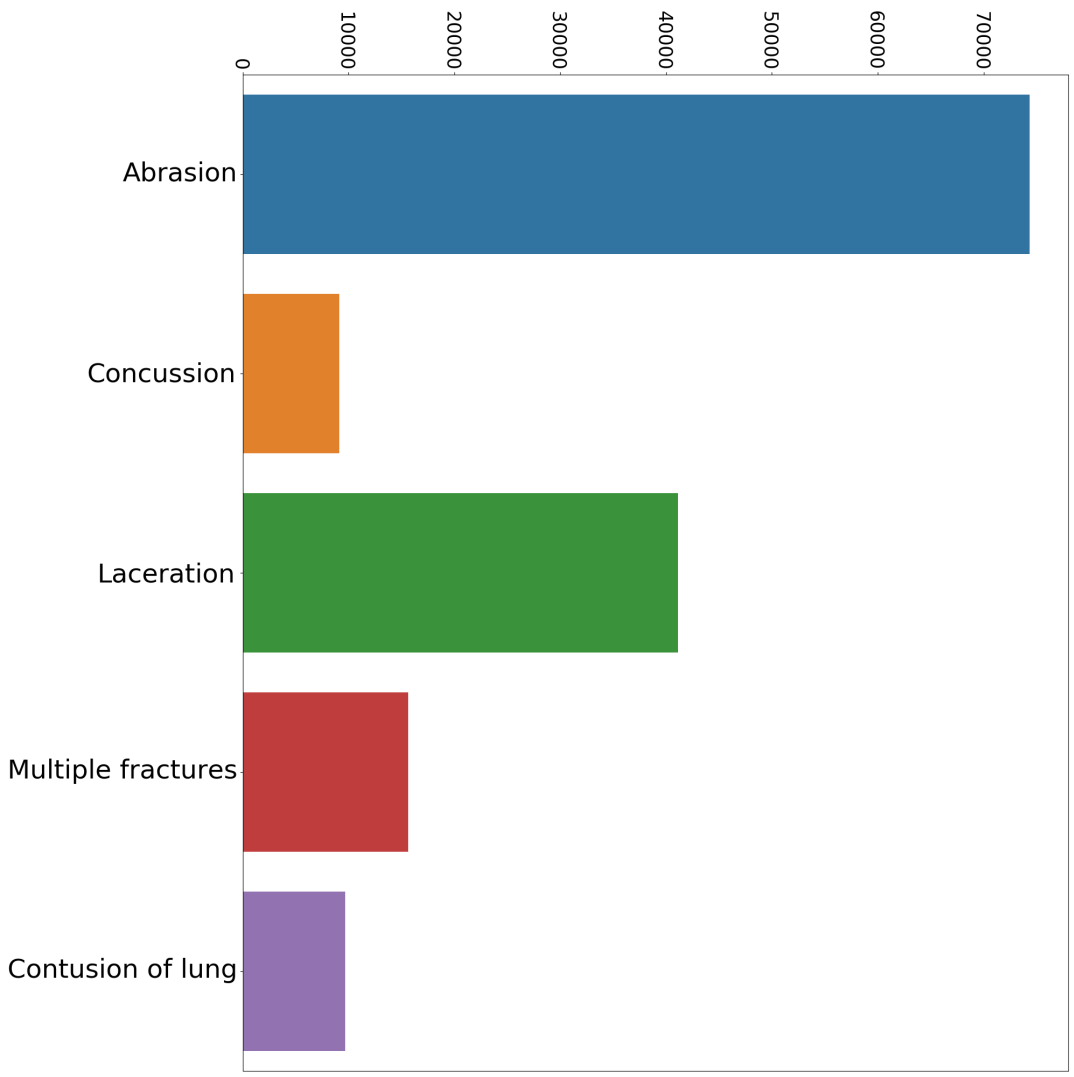


Figure 4.26.: Highest mortality traumas in motorcycle accidents

Furthermore, the age distribution for the five traumas chosen can be observed in Table 4.7. In general, all distributions show similar values for survivors and deceased patients, with the most different cases being in concussion trauma. The difference in the age distribution of concussion can be explained as concussion being a trauma that is not very dangerous unless under certain conditions older people maybe more prone to complications.

Moreover, in this trauma there was no data of deceased female patients, as shown in Figure 4.24. The distribution for the survivors can be seen in Figure 4.27. It is very interesting that no data of mortal concussion existed in the database, when looking at the distribution high probabilities of trauma survivors exist in women between 50-60 years old.

Abrasion				
	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
Male	38,59	15,24	49,33	16,27
Female	41,36	13,95	48,71	21,97
Concussion				
	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
Male	37,97	15,82	54,96	25,49
Female	73,8	38,45	-	-
Laceration				
	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
Male	38,64	15,17	42,95	16,81
Female	43,31	15,63	48,43	13,4
Multiple fractures				
	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
Male	45,36	15,44	48,93	16,65
Female	48,95	14,19	50,05	14,52
Lung contusion				
	Survived		Deceased	
	Mean	Standard deviation	Mean	Standard deviation
Male	37,4	15,44	43,91	16,66
Female	41,72	14,06	53,02	16,65

Table 4.7.: Age statistics in traumas from motorcycle accidents

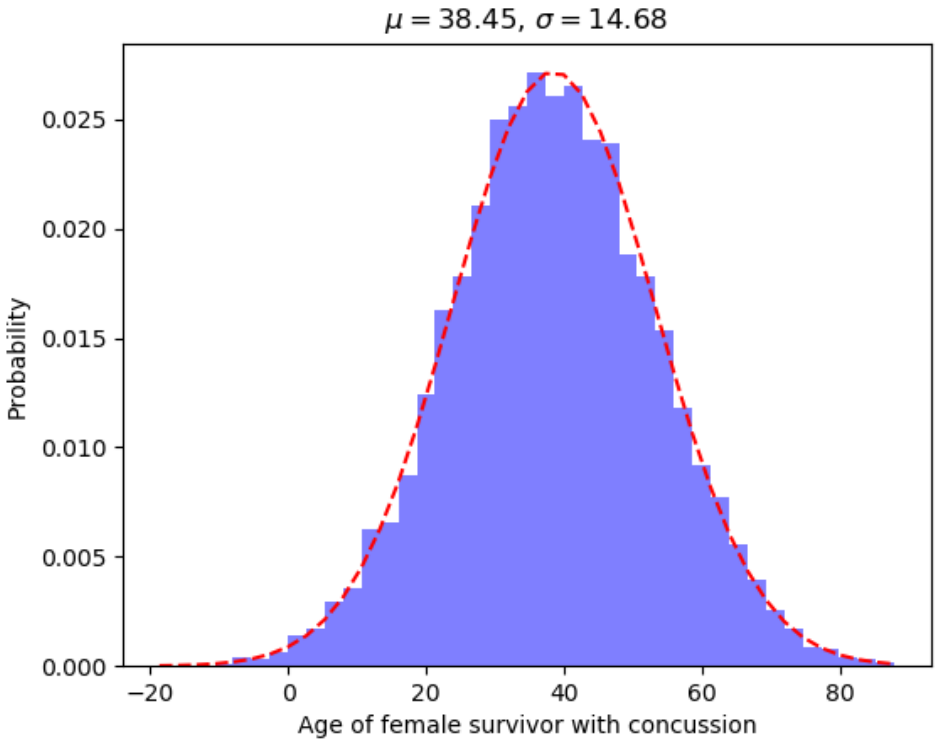
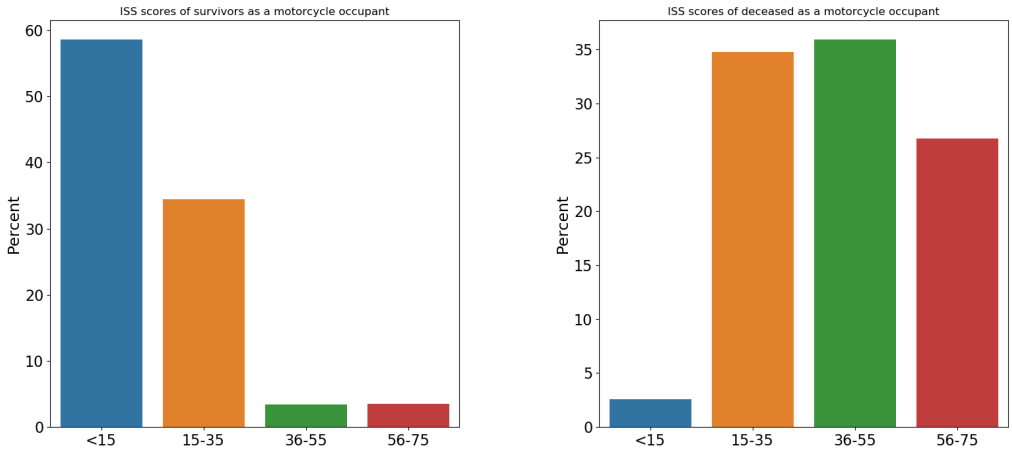


Figure 4.27.: Distribution of age in female patients with concussion in motorcycle accidents

4.1.7.2. Difference of trauma scores in motorcycle accidents

When analyzing the trauma scores for motorcycle accidents, ISS returned an out of the norm value, unlike the results shown in Subsection 4.1.2 in Figure 4.5b, which had a high number of patients with less than 15 of ISS score, and the predominant one was 15 to 35, considered a non-severe injury. Figure 4.28b shows the predominant one as 36 to 55, considered a severe injury, and the number of people with less than 15 of ISS score is significantly lower.

This may be caused because unlike some other traumas, in motorcycle accidents the hit is taken fully by the body, which means that the injuries can be seen more easily causing the initial evaluation to be worse than in other cases, which failed because of complications.



(a) ISS groups of survived patients **(b)** ISS groups of deceased patients

Figure 4.28.: ISS results of patients in motorcycle accidents

4.2. Correlation analysis

In this Section the analysis of linear correlations between the data is exposed. The aim of this research was to find a direct relation between variables that allowed us to develop a predictive model to foresee the evolution of trauma patients.

4.2.1. The process

In order to obtain the correlation between the different variables we had to join the respective tables together. As it was seen in Section 3.1, the database contains 32 different tables, each of them with its own information being linked to others by the *INC_KEY* label.

However, this primary key was not unique in all the tables, with tables such as ICD10_ECODE having up to 50 records per *INC_KEY*. This proved to be a problem when joining the tables together. The computer used to run the analysis had only 16 GB of RAM available, and when merging the whole database as a single table the number of records where over 6 millions, amounting to a total of 15,2 GB of RAM used. As that amount of RAM was used only as storage, when the calculations started it capped out and failed the process. To solve this problem, two solutions were implemented:

Data reduction: With this solution we only kept the first record of every patient, as the record was not chosen manually the data stayed untainted. However, a huge loss of data occurred, although the table still had around a million records, 5 millions records were lost. Due to the fact, this method was discarded as viable and was only used to show the correlation of all the tables merged (Figure 4.29),

as the data used in it was smaller than in the subgroup cases, this figure was not analyzed and is only to show a general view of the relations.

Subgroups of tables: This method consisted in analyzing the correlation using only two tables at a time, reducing drastically the memory cost of the load. This way we would still be able to test the correlations, but in a much smaller scale, without the need to cut down records of the patients.

It is also worth mentioning that as the data was stored in a CSV file, the types of the columns were selected by Pandas and a lot of columns were using nonoptimal types. This was mainly the case with number columns as they were categorized as string columns. To fix this, a function was implemented to check for wrong types. However, when this process was finished, another problem was found. The number types were assigned to float64 types when they only contained integers. The problem resided in the missing data as Pandas treated NaN values as float, and thus this increased the size of the database in the memory. To solve this problem a specific integer was assigned to NaN values, such as -1. This reduced the size of the object in the memory to that of an int16 or even int8 in some cases (as shown in Table 4.8).

Type	Size [Bytes]
int8	1
int16	2
int32	4
int64	8
float16	2
float32	4
float64	8
string	Number of characters * size of int8

Table 4.8.: Type size reference

The same problem occurred with the strings, when loaded a categorical type was assigned. As most of the strings were not unique, for example, gender column only contained 3 distinct values: Male, Female, and Not Known/Recorded; they could be substituted by a code (Male: 1, Female: 2, Not Known/Recorded: 0). This will support the correlation analysis that will be performed, as the strings have to be converted to numeric values in order for the correlation method to work.

By applying these methods, the memory consumption of the database was successfully reduced. Although, the merged database could not be loaded, the size of the table groups was dramatically reduced. Some of the tables occupied nearly 2 GB of RAM, and after the conversion used only 200 MB, a tenth of their original memory use.

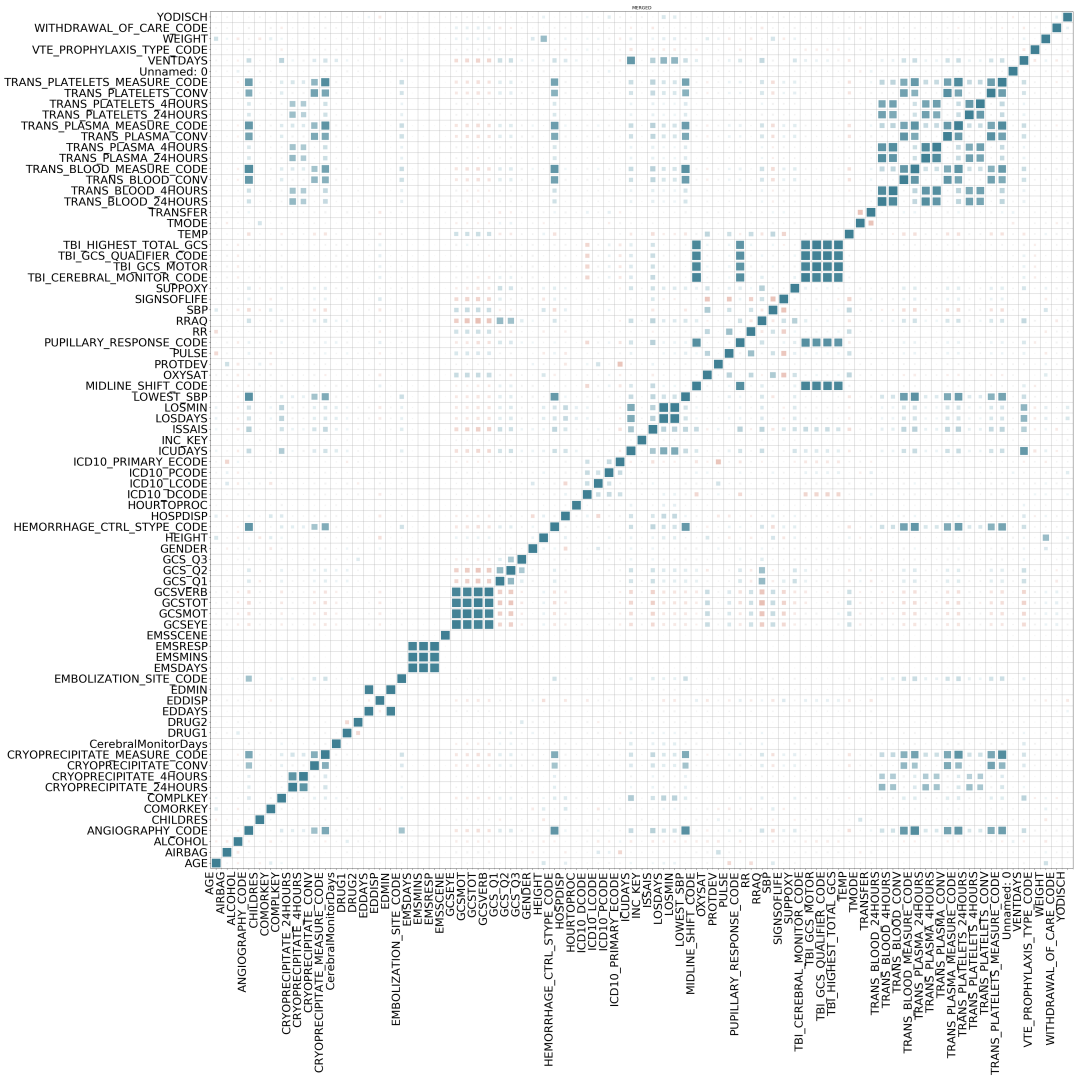


Figure 4.29.: Correlation of merged database

4.2.2. The analysis

With the initial preprocessing done, we were ready to move to the correlation analysis. However, some further measures/actions need to be considered before moving to this analysis.

Firstly, we needed to decide which tables to cross in the correlation analysis. To do so, the description tables ending with *DES* were discarded. Those tables only contain the description of the numeric codes of their respective tables, which link to the patients records with a code. An example of this can be seen in Table 4.9, where columns *ICD10_DCODE* in both tables are the same, but one contains the reference to the patient(*INC_KEY*), and the other contains the description of the code (*ICD10_DCODEDES*).

Contents of ICD10_DCODE		Contents of ICD10_DCODEDES	
INC_KEY	ICD10_DCODE	ICD10_DCODE	ICD10_DCODEDES
160594005	Z99.89	Z99.89	Dependence on other enabling machines and devices
160013721	Z98.89	Z98.89	Other specified postprocedural states
160256005	Z97.2	Z97.2	Presence of dental prosthetic device (complete) (partial)
160611414	Z96.659	Z96.659	Presence of unspecified artificial knee joint

Table 4.9.: Sample of ICD10_DCODE and ICD10_DCODEDES tables

After discarding the description tables we chose the rest of the tables based on the information they contain, and the possible relation between each other. The groups of tables analyzed:

AISPCODE and DISCHARGE: The reason for this analysis was to find a relation between the degree of severity of the injury and the discharge information of the patient.

AISPCODE, DISCHARGE, ICD10_ECODE, ICD10_DCODE, and ICD10_LOC: In order to expand the previous analysis, the discharge information of the patient was considered important to correlate them with not only the degree of severity of the injury, but also with the external cause of injury, the diagnosis code and the location code.

DEMO and COMORBID: With this pair, we wanted to analyze if there was any relation between the demographic information of the patients and the comorbid conditions of patients upon arrival in the hospital.

DEMO and DCODE: We wanted to see if there was a relation between the demographic information of the patients and the diagnosis code of the injury.

DEMO and DISCHARGE: Correlating these two tables, the demographic information of the patient will be analyzed with respect to the discharge information which provides details with respect to the length of the stay in the different facilities of the hospital.

DEMO and ECODE: Same case that DEMO and DCODE, but with the external cause of injury contained in table ECODE.

DEMO and ICD10_DCODE: We wanted to see if there was any difference between ICD-9 diagnosis code of injury and the ICD-10 diagnosis code of injury.

DEMO and ICD10_ECODE: Same case as in the correlation of DEMO and ICD10_DCODE, but in this case with the external case of injury codes.

DEMO and VITALS: With this pair we were searching for a possible relation between the demography and the vitals of the patient.

DISCHARGE and COMORBID: We wanted to see whether how much affected the situation of comorbidities was represented in liner correlation.

ED and DISCHARGE: We wanted to see if by being two related events, one being the Emergency Department, and the other being the discharge information, some correlation could be seen and used.

TRANSPORT and DISCHARGE: See if there was a relation between the type of transport and the discharge status.

VITALS and DISCHARGE: For this one we wanted to see if there was a huge relation between patient vitals and their discharge.

Secondly, the Pearson coefficient was used for the correlation analysis. This was chosen due to the disposition of the database. With the variables tending more towards a linear function than towards a curve function. Moreover, there was a large number of outliers found, these were removed from the sets. The method used to remove the outliers was Z-score for its simplicity and accuracy.

Finally, the rows containing missing data and typing errors were filled with an arbitrary number. In some cases, when the column had a huge number of missing data the column was dropped from the row and not analyzed. As the volume of data was already big enough, the rows dropped did not affect noticeably the results.

5. Experimentation

In this Chapter, we are going to expose and analyze the results from the correlation analysis will be presented.

5.1. Correlation results

First of all, the format used in the correlation tables is explained. The results are color coded: opaque blue indicates perfect positive correlation (1), opaque red indicates perfect negative correlation (-1). The closer the value is to 0, the smaller the square is, and the transparency increases. An example of this can be seen in Figure 5.1 and Table 5.1.

c	-1	0.5	1
b	-0.5	1	0.5
a	1	-0.5	-1
	a	**b**	**c**

Table 5.1.: Table of correlation sample

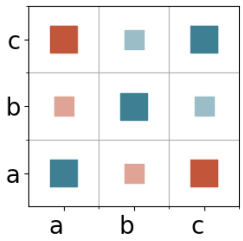


Figure 5.1.: Figure of correlation sample

AISPCODE and DISCHARGE

The first correlation analysis is between tables AISPCODE and DISCHARGE (Figure 5.2). In this case the highest correlations were between *ICUDAYS-LOSDAYS*, *ICUDAYS-LOSMIN*, *LOSMIN-LOSDAYS*, and *ICUDAYS-VENTDAYS*:

- *ICUDAYS* represents the time spent in the ICU.
- *LOSDAYS* represents the total time spent in the hospital.
- *VENTDAYS* the time spent in a ventilator.

AISPCODE columns are *SEVERITY* and *PREDOT* which provide information about the severity codes and the codes that reflect the part of the body affected by the trauma injury. None of them present any correlation with the DISCHARGE variables. However, the correlations found in DISCHARGE are rather interesting.

On the one hand, we have *ICUDAYS-LOSDAYS*, which tells us that most of the patients that suffer a trauma end up in the ICU. Moreover, *ICUDAYS-VENTDAYS* confirms the data found in Subsection 4.1.4, where in Figure 4.13 a strong correlation between the two columns can be seen, and the studies found [73], that most of the ICU patients are mechanically ventilated; therefore, the strong correlation found confirms this fact.

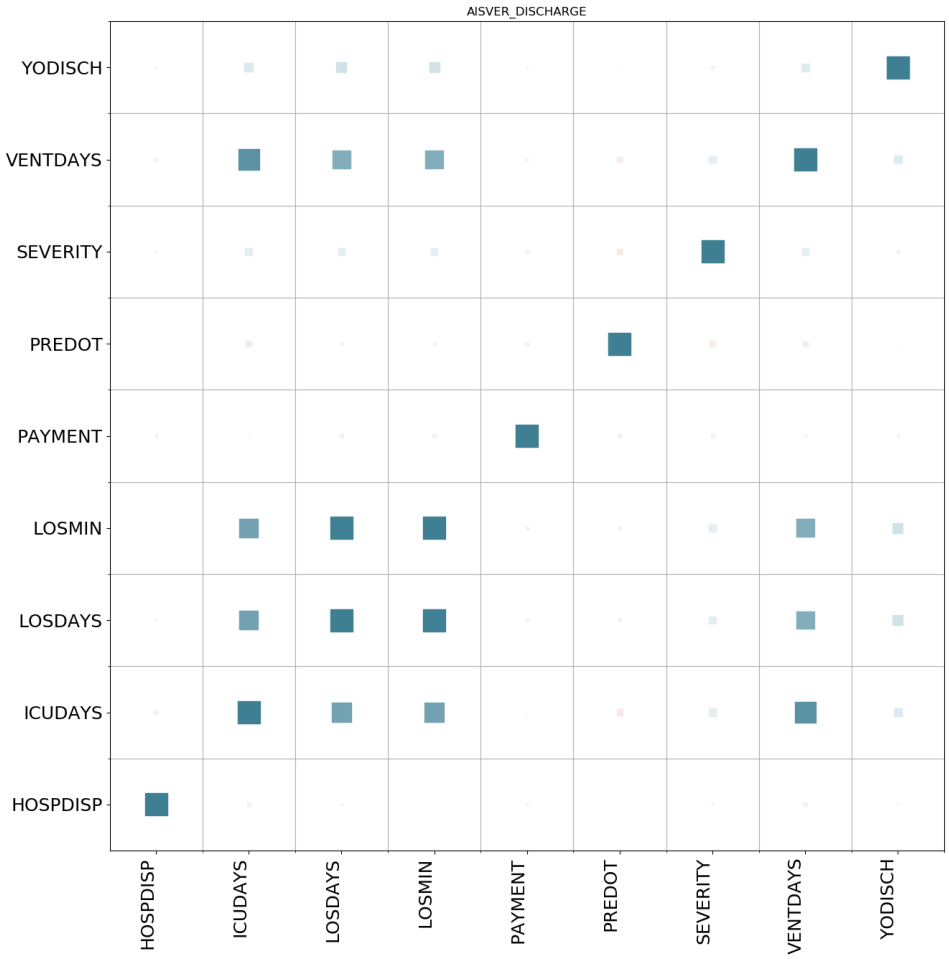


Figure 5.2.: AISPCODE and DISCHARGE correlation

AISPCODE, DISCHARGE, ICD10_ECODE, ICD10_DCODE, and ICD10_LOC

The second correlation analysis is between AISPCODE, DISCHARGE, ICD10_ECODE, ICD10_DCODE, and ICD10_LOC. As we can see in Figure 5.3 the correlation between these tables is rather poor, with the exception of *Trauma_Type-ICD10_ECODE*, *Trauma_Type-Intent*, and*Intent-ICD10_ECODE*.

- *Trauma_Type* establishes whether a trauma is blunt, penetrating, or a burn. From table ICD10_ECODE.
- *ICD10_ECODE* is the code for the external cause of injury. From table ICD10_ECODE.
- *Intent* is whether the injury was self-inflicted, unintentional, or intentional. From table ICD10_ECODE.

As we can see, those three columns pertain to same the same table and as such

they were expected to be directly related.

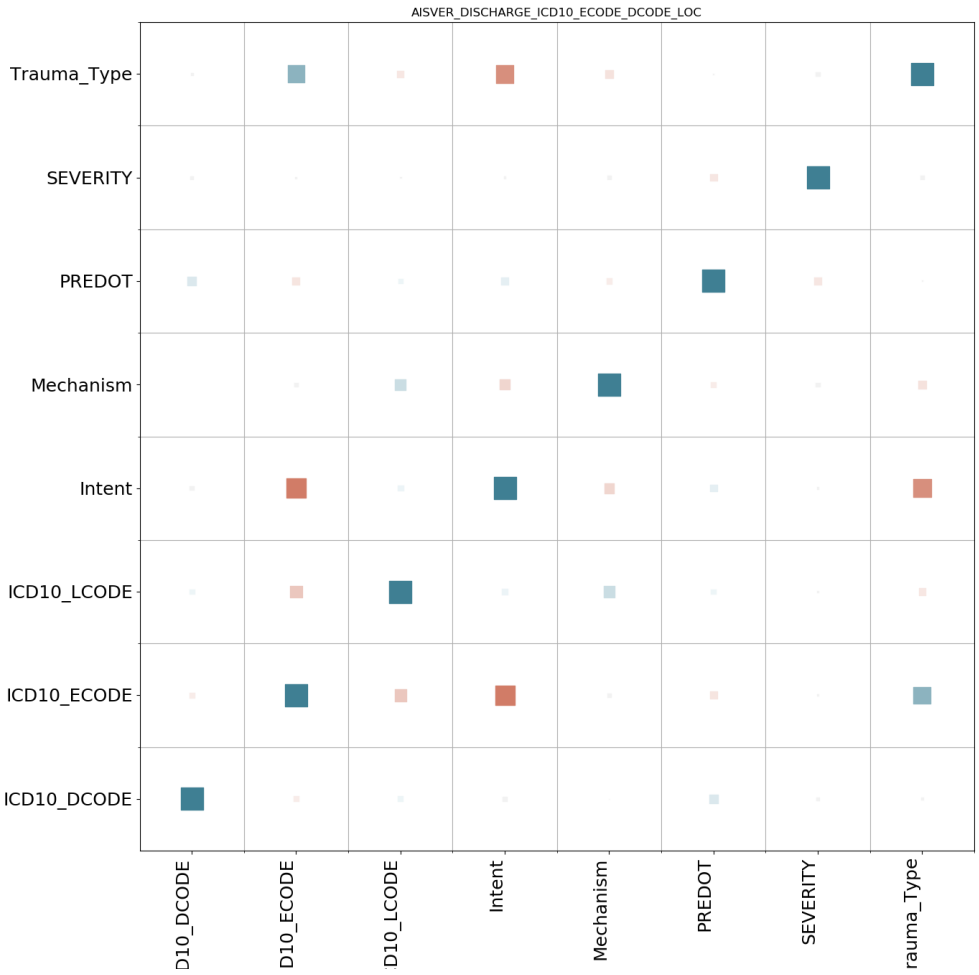


Figure 5.3.: AISPCODE, DISCHARGE, ICD10_ECODE, ICD10_DCODE, and ICD10_LOC correlation

DEMO and COMORBID

The third analysis is focused on the DEMO and the COMORBID tables (Figure 5.4). In this case the only positive correlation is the case of *YOBIRTH-AGE*:

- *YOBIRTH* is the year of birth of the patient.
- *AGE* is the age of the patient.

As can be seen form the data these columns contain the correlation is expected to be strong. However, here we can see how the missing data affects the results, although in a smaller way, as the correlation between *YOBIRTH* and *AGE* should be 1, but is 0,79. This is because *YOBIRTH* has missing data, causing the correlation to be lower than expected, as the value in *YOBIRTH* is -1 (the NaN value).

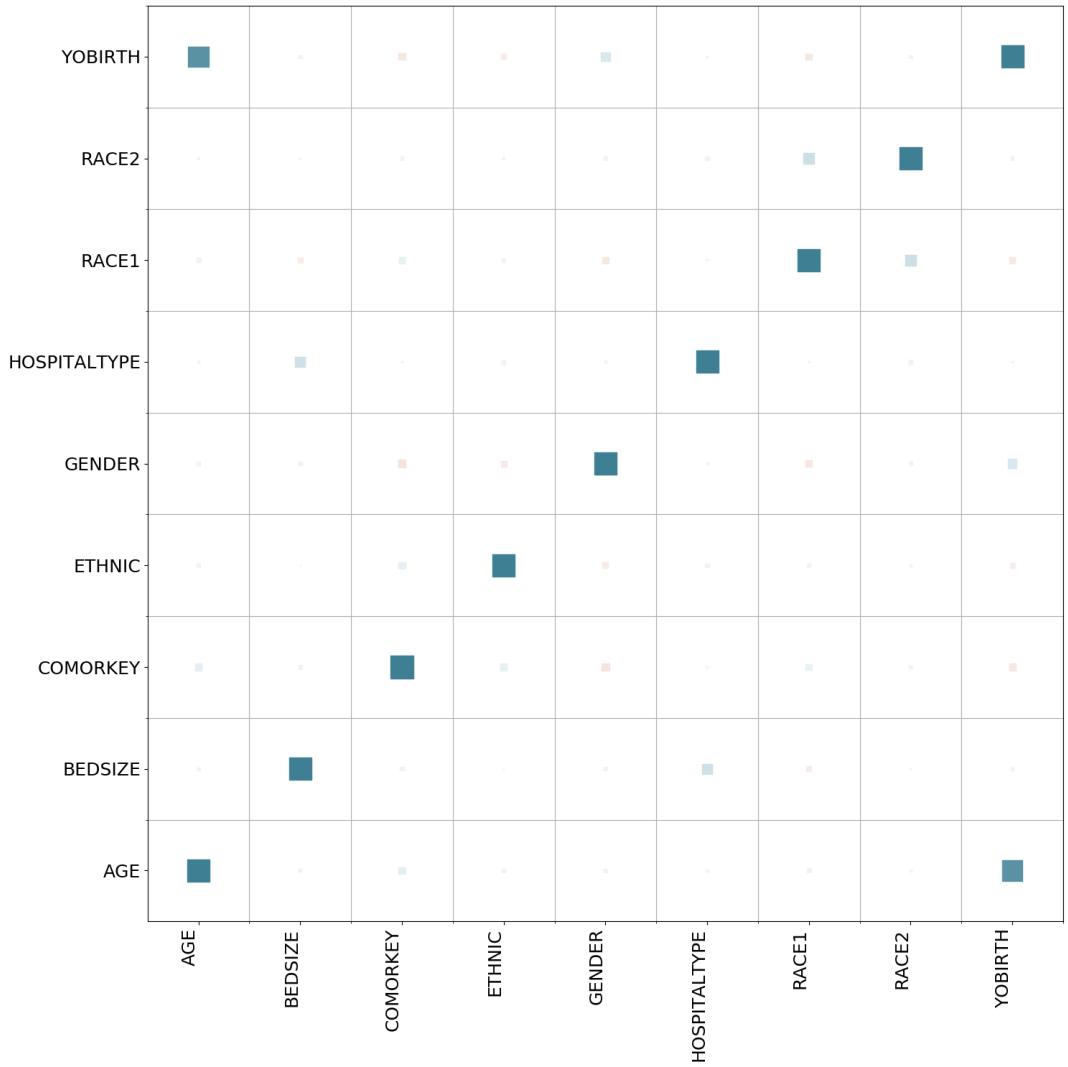


Figure 5.4.: DEMO and COMORBID correlation

DEMO and DCODE

In the case of the tables DEMO and DCODE, the correlation between the two tables is almost null, as we can see in Figure 5.5. The column from DCODE is *DCODE*. The diagnosis code informs of the injury suffered by the patient, a sample of the table can be seen in Table 5.2. However, it does not have any correlation neither positive nor negative with the DEMO columns.

DCODE	Description
372.72	Conjunctival hemorrhage
372.73	Conjunctival edema
374.34	Blepharochalasis
376.30	Exophthalmos unspecified
376.32	Orbital hemorrhage
376.33	Orbital edema or congestion

Table 5.2.: Sample of diagnosis codes

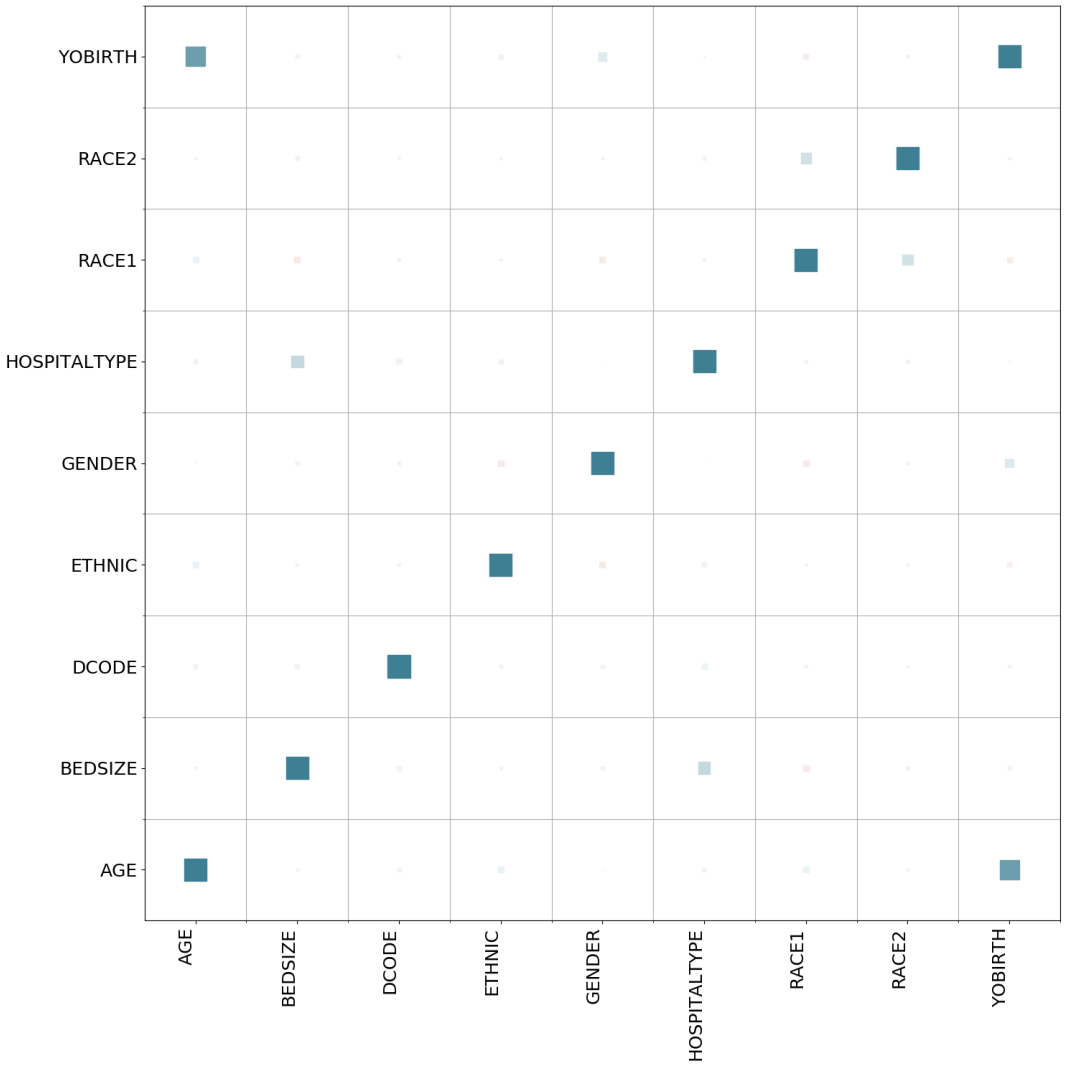


Figure 5.5.: DEMO and DCODE correlation

DEMO and DISCHARGE

Although expected to find a possible relation between the demographic information of patients, specially between the age and discharge status, nothing significant was found, as seen in Figure 5.6. The only correlations visible are the ones

already analyzed from within table DISCHARGE (*ICUDAYS-LOSDAYS*, *ICUDAYS-LOSMIN*, *LOSMIN-LOSDAYS*, and *ICUDAYS-VENTDAYS*).

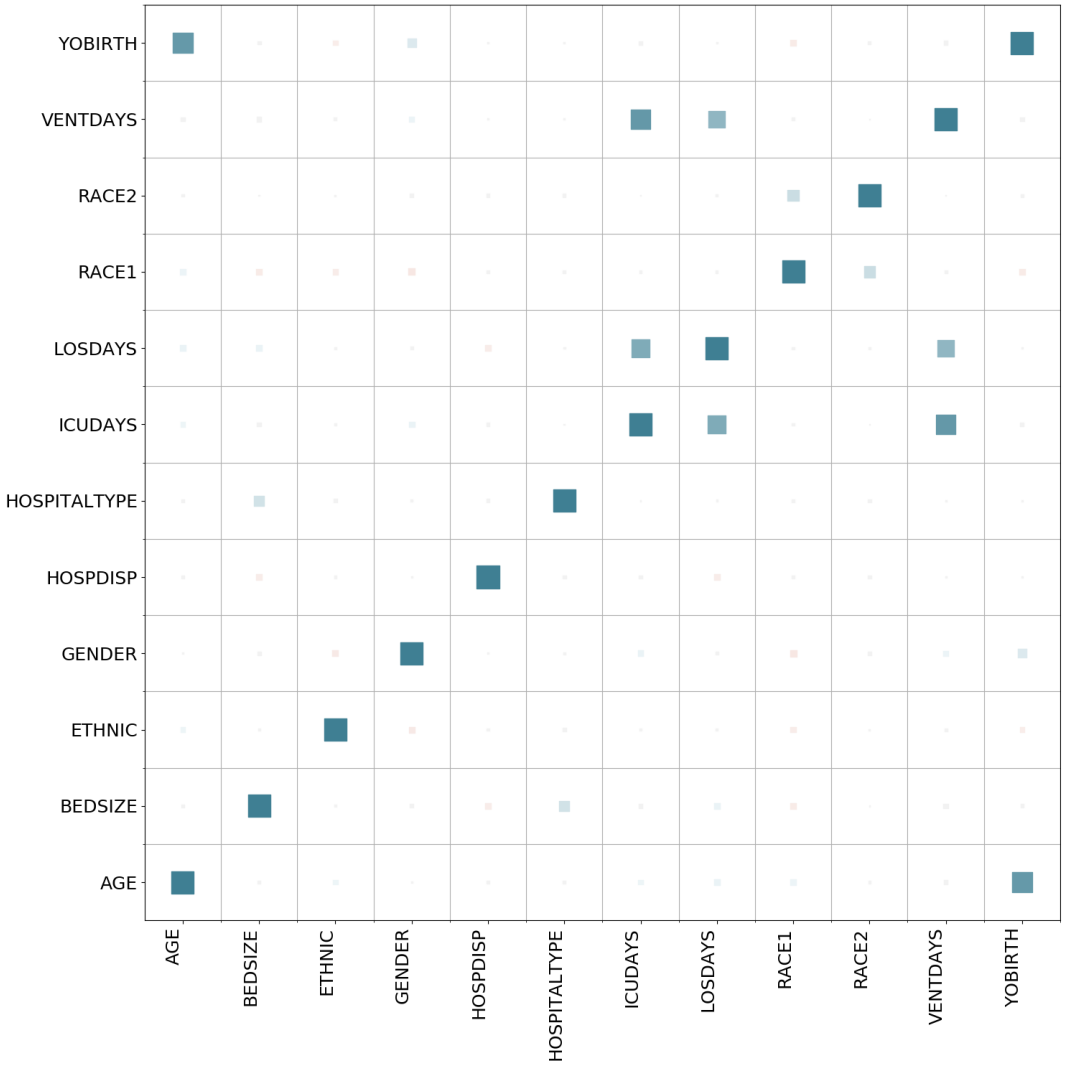


Figure 5.6.: DEMO and DISCHARGE correlation

DEMO and ECODE

The sixth correlation analysis was done between the DEMO and the ECODE tables (Figure 5.7). With this pair of tables we found positive correlations in *MECHANISM-ECODE*, *INTENT-ECODE*, *INJTYPE-ECODE*, and to a lesser extent *MECHANISM-INTENT*.

- *INJTYPE* defines the type of trauma, wether is blunt, penetrating, or burn.
- *INTENT* is if the trauma was caused consciously or unconsciously.
- *MECHANISM* is how the trauma took place, for example, as a pedestrian or by drowning.

Although at first this looks like positive it isn't very surprising, as again those columns are all part of the same table (ECODE), meaning that there is no correlation between tables DEMO and ECODE. Moreover, as those columns are inherently related, is to be expected that a positive correlation exists. In Table 5.3 a sample of how the external cause of injury table is stored in the database (the result of joining ECODE and ECODEDES) can be seen.

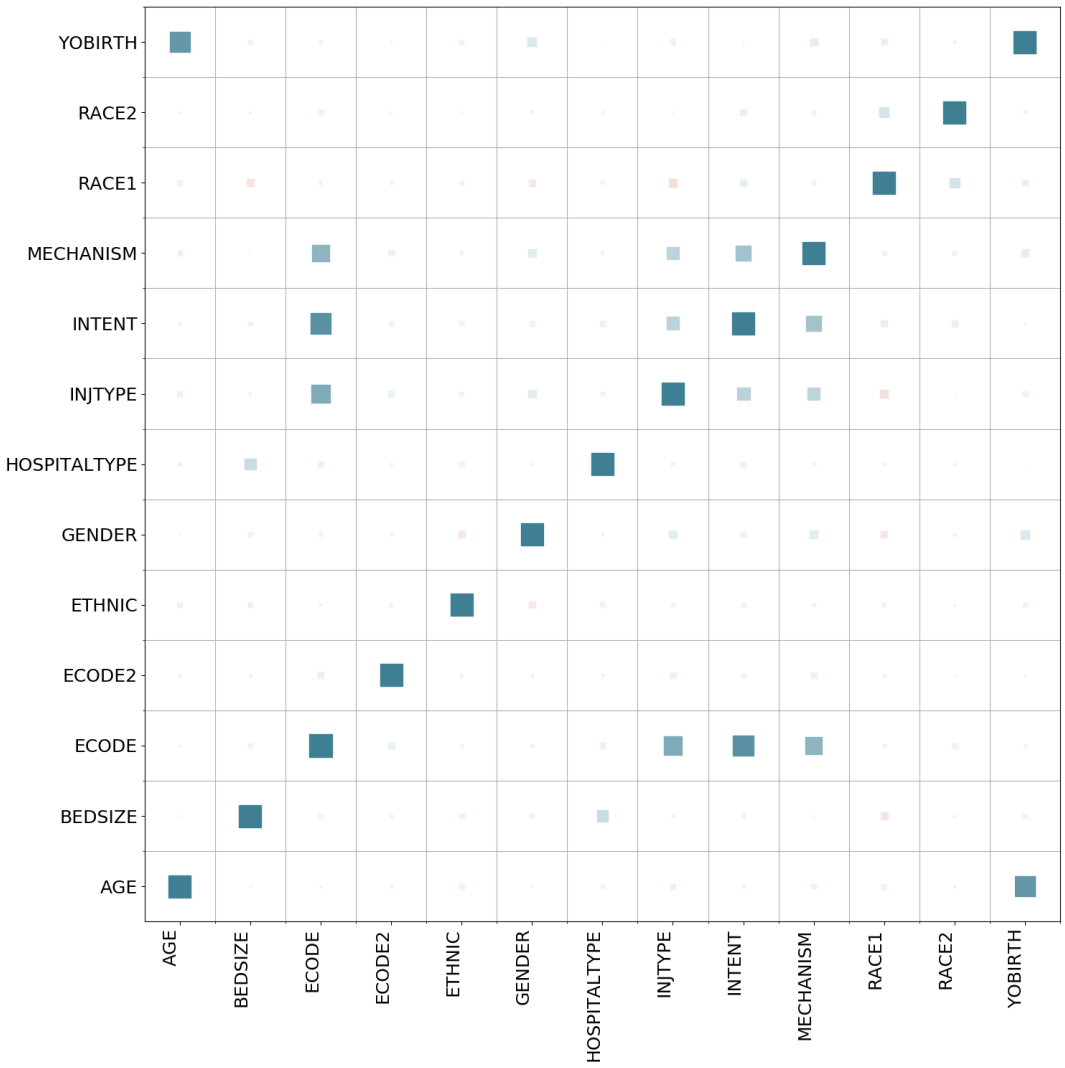


Figure 5.7.: DEMO and ECODE correlation

DEMO and ICD10_DCODE

In the case of the DEMO and the ICD10_DCODE tables, no correlation was found. The column from the ICD10_DCODE table is *ICD10_DCODE* and its correlation with the other variables is almost 0 in all cases, as seen in Figure 5.8

ECODE	Description	Type of injury	Intent	Mechanism
805.2	Railway, Hit by Rolling Stock - Pedestrian	Blunt	Unintentional	Pedestrian, other
806.3	Oth Spec Railway Accident - Pedal Cyclist	Blunt	Unintentional	Pedal cyclist, other
815.2	Oth MVA Traffic, Highway Collision - Motorcyclist	Blunt	Unintentional	MVT Motorcyclist
983.0	Hang/Strangle/Suffocate, Un/Intentional- Hanging	Other/unspecified	Undetermined	Suffocation
985.0	Firearms/Explosives, Un/Intentional - Handgun	Penetrating	Undetermined	Firearm
987.0	Fall From High Place, Un/Intentional - Residential Premises	Blunt	Undetermined	Fall

Table 5.3.: Sample of the external case of injury

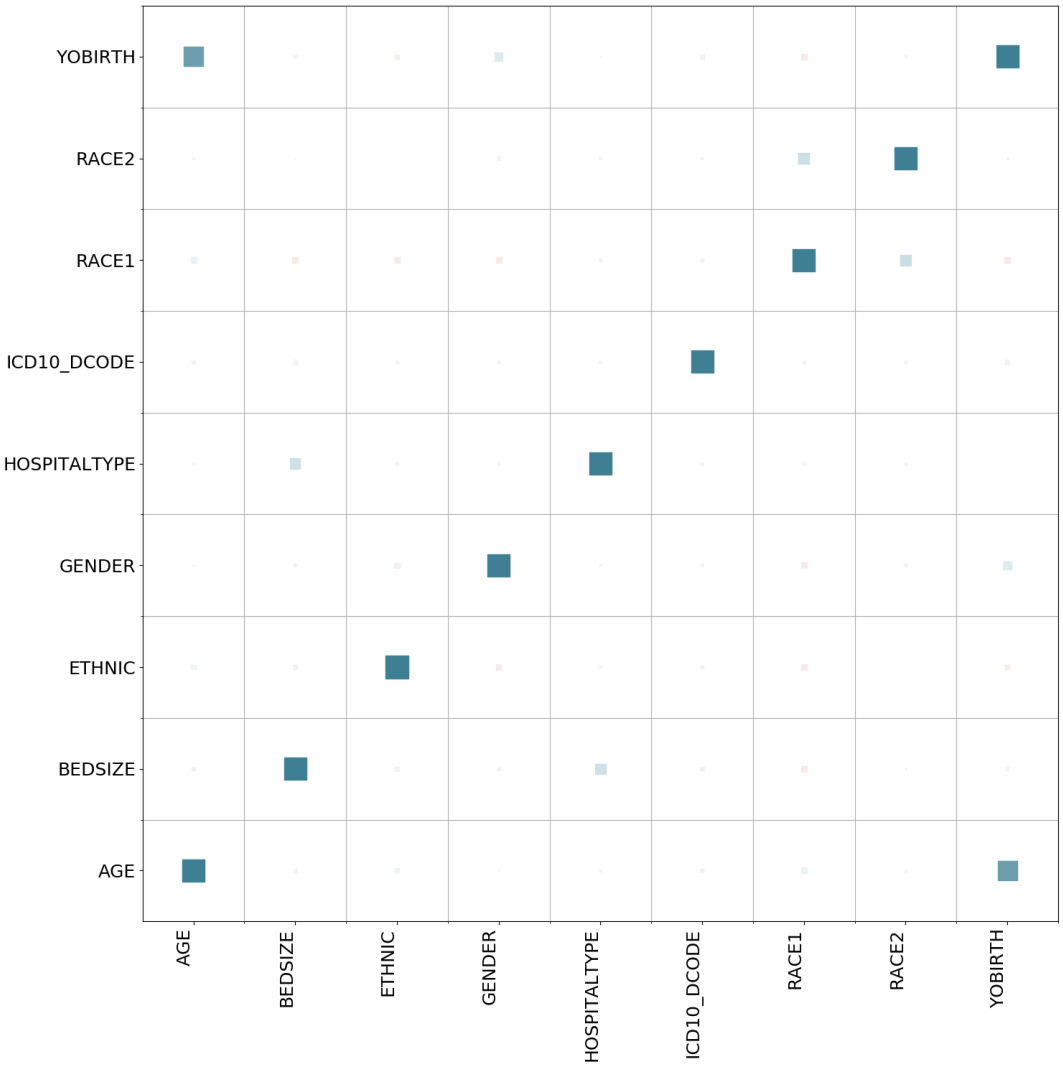


Figure 5.8.: DEMO and ICD10_DCODE correlation

DEMO and ICD10_ECODE

The eighth correlation analysis between the DEMO table and the ICD10_ECODE table (Figure 5.9) proves to be as unfruitful as the previous correlation showing no correlations between the two tables. The only correlation outside of what has already been analyzed in the previous correlation analyzes is *INTENT-ICD10_ECODE*, which is to be expected as both columns belong to the same table, ICD10_ECODE, indicating that they are inherently correlated.

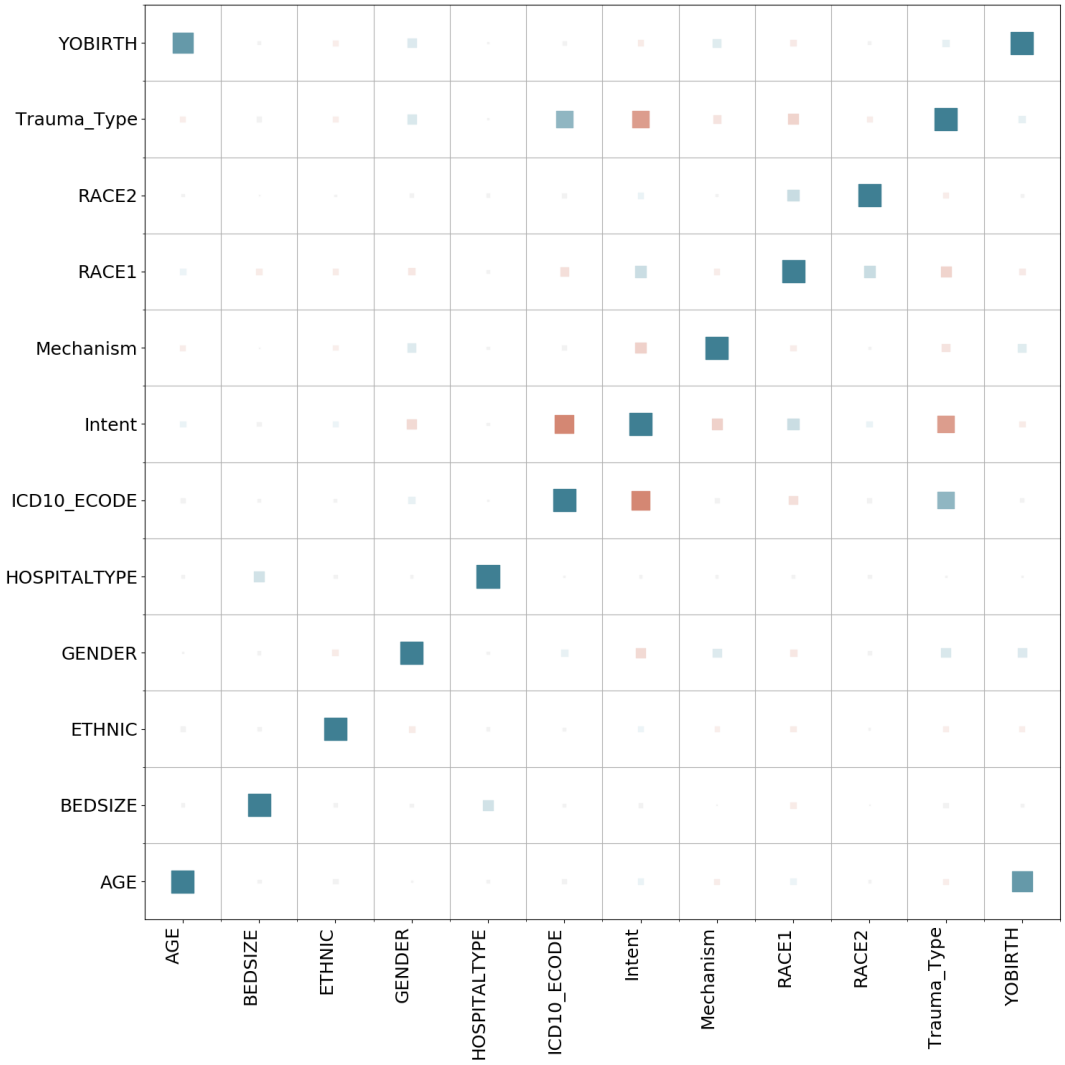


Figure 5.9.: DEMO and ICD10_ECODE correlation

DEMO and VITALS

The ninth correlation is that of the DEMO and the VITALS tables. As we can see in Figure 5.10 there are several positive correlations. with this relation. There are important correlations between then GCS variables: the verbal Glasgow Coma Score (*GCSVERB*), the motor one (*GCSMOT*), the visual one (*GCSEYE*), and the

total Glasgow Coma Score (*GCSTOT*). This is due to the fact that the total score is a combination of the results obtained from the individual Glasgow Score values as explained in Subsection 2.1.1. Furthermore, there is another interesting group of correlations, which are the main vital constants of the patients together with the respiratory assistance assessment qualifier (*RRAQ*).

In Figure 5.10 can be seen is that the GCS correlates to the vital parameters of the patients with a seemingly strong correlation, confirming that the GCS, a priori, shows the condition of the patient's vitals. Moreover, those parameters also correlate among them, confirming also that there is a relation between vitals parameters. That phenomena could be seen also in Subsection 4.1.4 where we saw that all parameters reacted similarly following the state of the patient.

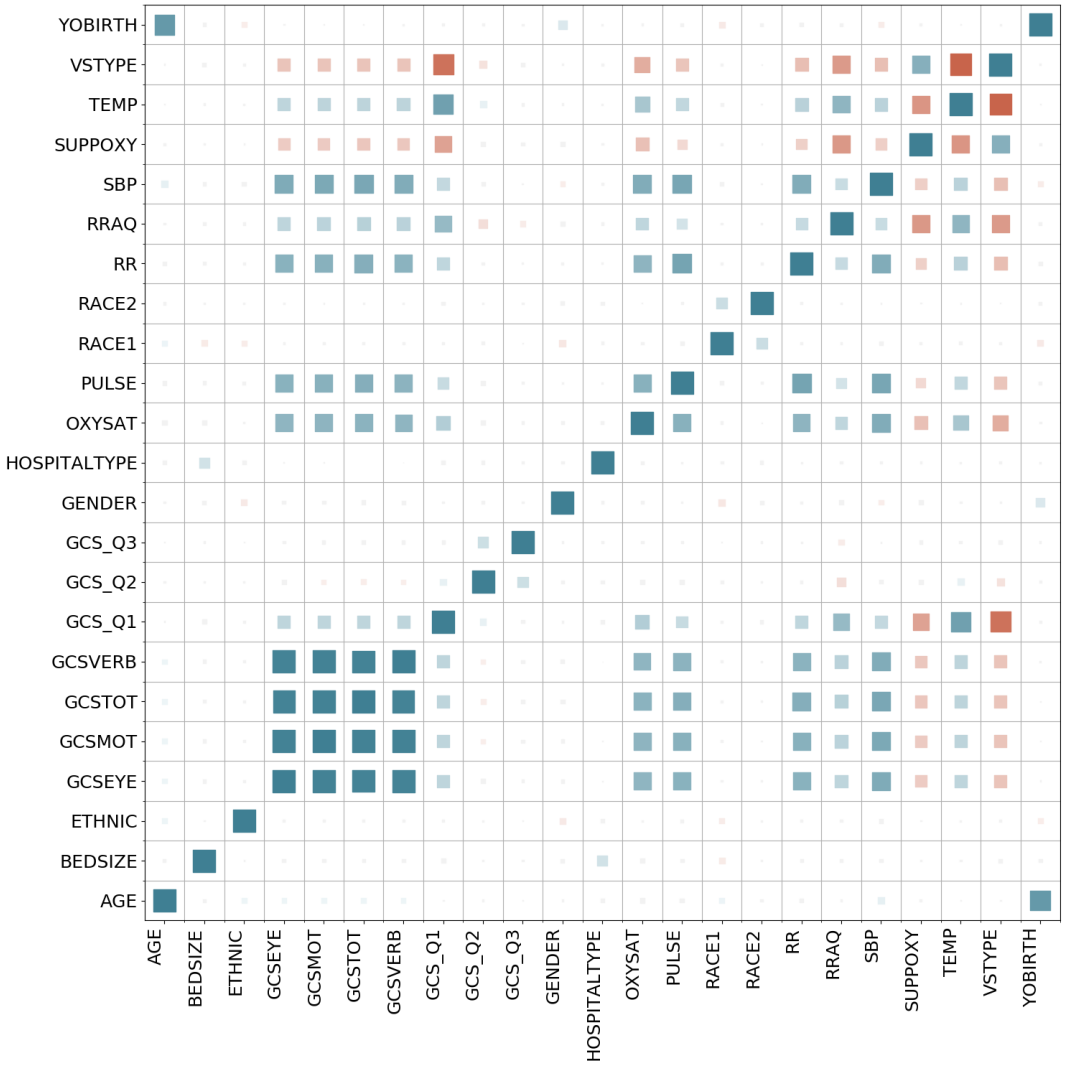


Figure 5.10.: DEMO and VITALS correlation

DISCHARGE and COMORBID

The next pair of tables to analyze are the DISCHARGE and the COMORBID tables together. As observed in Figure 5.11, based on the fact that the only column from COMORBID is *COMORKEY*. *COMORKEY* is the code of the additional conditions suffered by the patient, a sample of this column can be seen in Table 5.4. The correlation between the two tables is nonexistent.

COMORKEY	Description
11	Diabetes mellitus
28	Drug use disorder
19	Hypertension requiring medication
27	Mayor psychiatric illness

Table 5.4.: Sample of comorbid conditions in the database

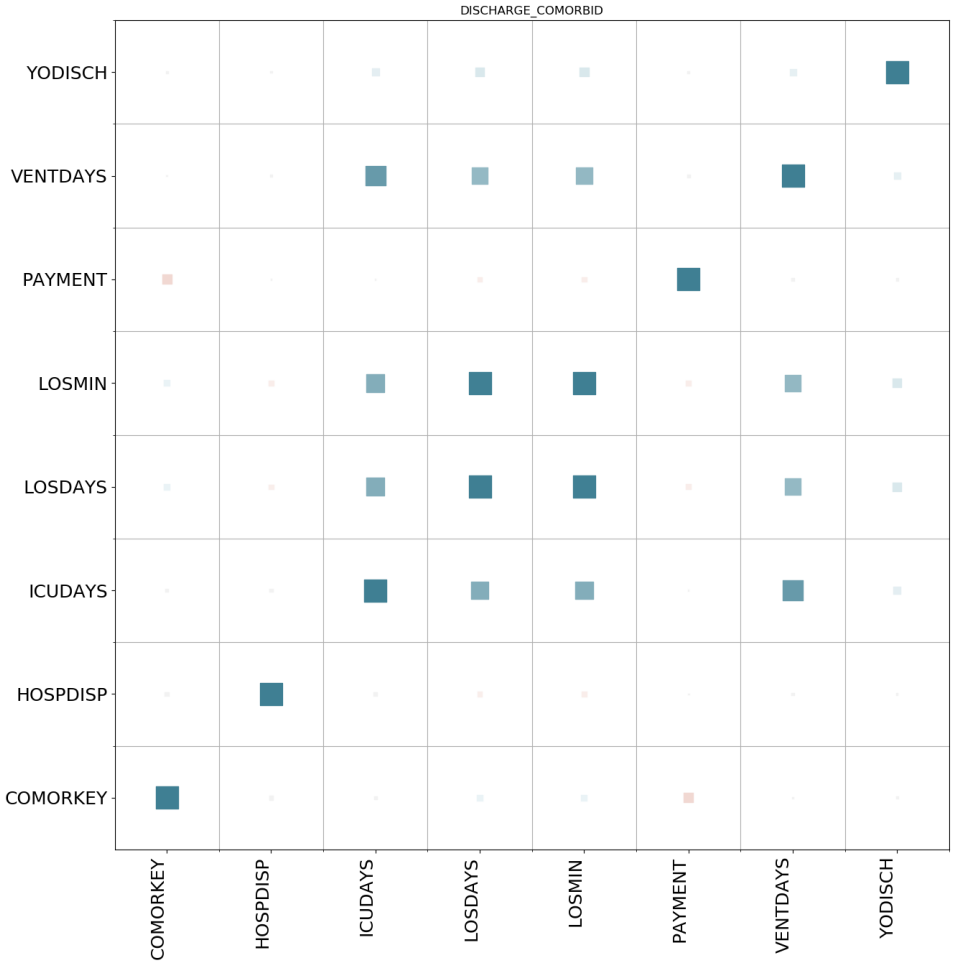


Figure 5.11.: DISCHARGE and COMORBID correlation

ED and DISCHARGE

The eleventh pair to be analyzed are the ED table and the DISCHARGE table. In Figure 5.12 we can observe the relation between *EMSRESP*, *EMSMINS*, and *EMSDAYS*, as well as *LOCATION* and *LECODE*.

- *EMSRESP* is the time elapsed to arrive to the scene in minutes.
- *EMSMINS* is the time elapsed to arrive to the hospital in minutes.
- *EMSDAYS* is the time elapsed to arrive to the hospital in days.
- *LOCATION* is the location where the injury happened.
- *LECODE* is the code of the location where the injury happened.

The reason for the relation of the fist group is caused by the fact that these three variables represent the same information, with the only differences that in one case the time unit of the variable is expressed minutes, and the other one is expressed in days, and that *EMSRESP* is the time elapsed to arrive to the scene in minutes, and *EMSMINS* is the time elapsed to arrive to the hospital.

In the case of *LECODE* and *LOCATION*, as *LECODE* is the code of the location where the injury was received, while *LOCATION* is the description of that code, a strong correlation is expected. A sample of this relation can be seen in Table 5.5.

LECODE	Location
0	Home
5	Street
1	Farm
6	Public building
7	Residential institution

Table 5.5.: Sample of the relation between *LECODE* and *LOCATION* columns

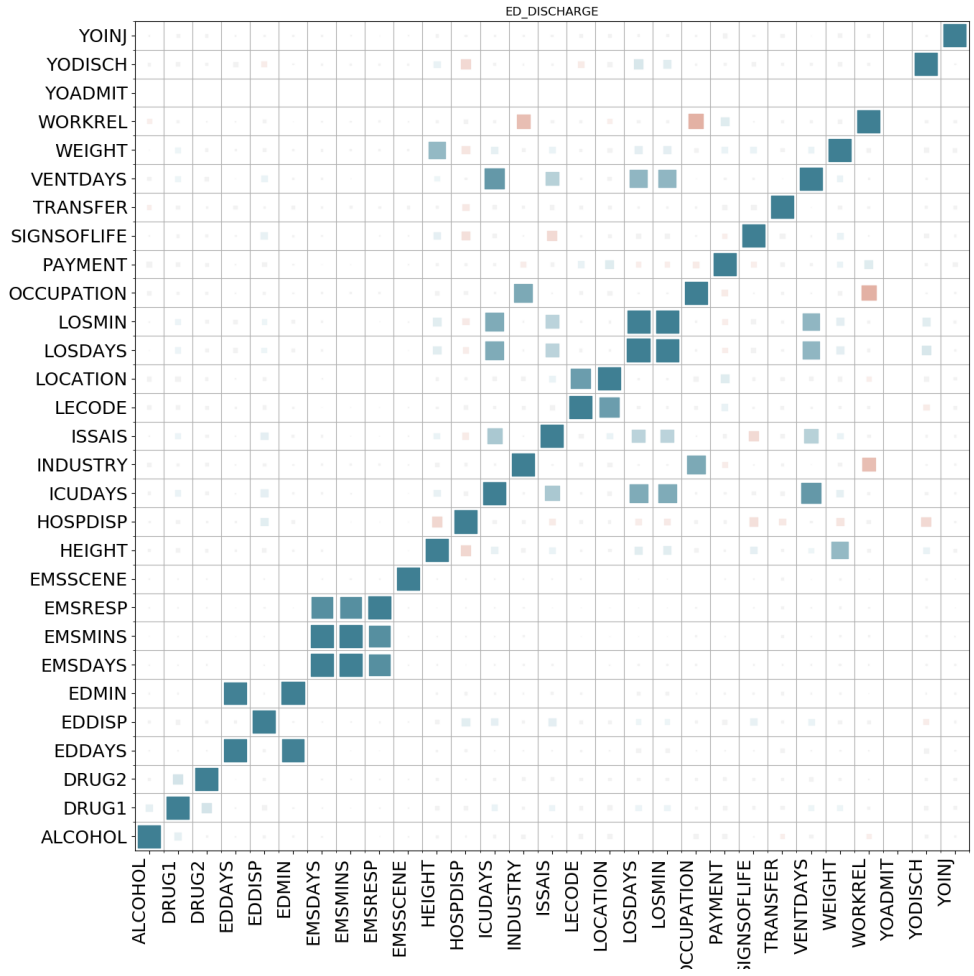


Figure 5.12.: ED and DISCHARGE correlation

TRANSPORT and DISCHARGE

The next pair of tables analyzed are the TRANSPORT and DISCHARGE tables. In Figure 5.13 we can see how there is no correlation between the two tables, as the columns from TRANSPORT are *TMODE* and *TRANTYPE*, and have neither positive, nor negative correlation with columns from DISCHARGE.

- *TRANTYPE* is the priority of the transport, it has two possible values, ”Other” and ”Primary”.
- *TMODE* is the mode of transport, ambulance, helicopter, police, or private (by car or walking).

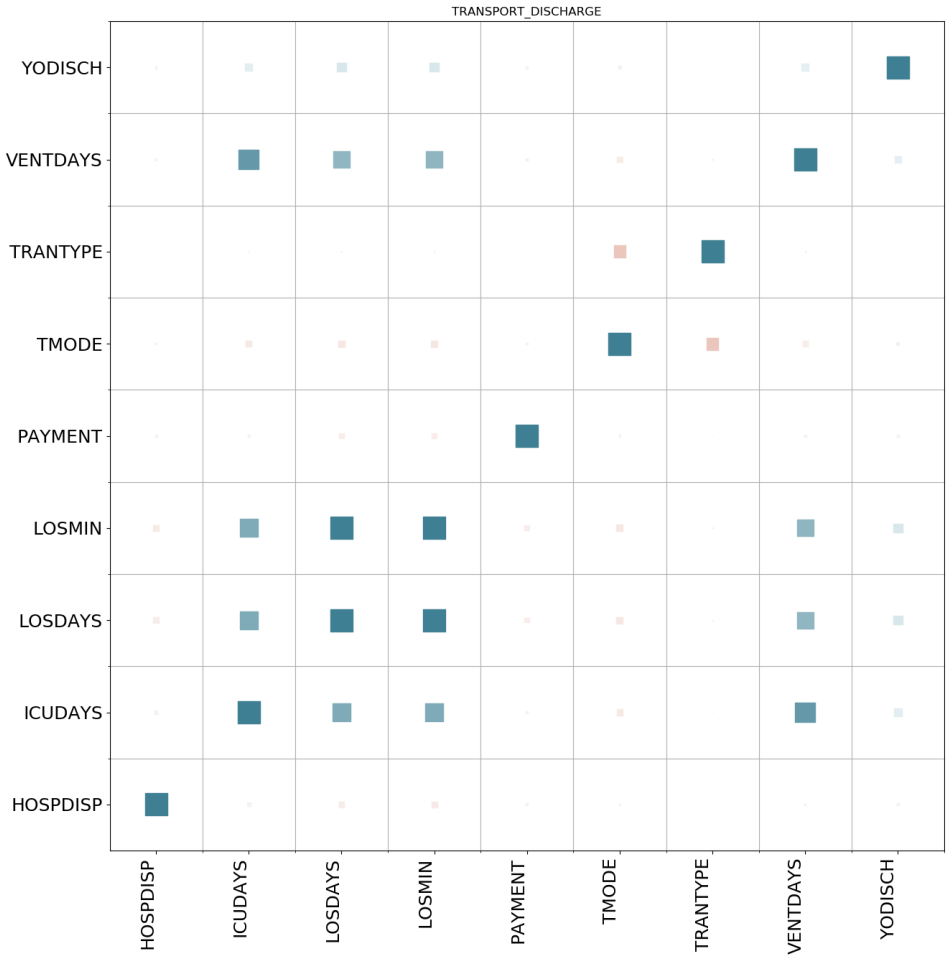


Figure 5.13.: TRANSPORT and DISCHARGE correlation

VITALS and DISCHARGE

The final correlation analyzed is between the VITALS and DISCHARGE tables; however as seen in Figure 5.14, apart from those correlation already analyzed nothing new is observed. Furthermore, as the only column in the Figure from the DISCHARGE table is *HOSPDISP* (this column specifies whether the patient survived or not) we can see that its correlation with other columns is 0, or very close to 0.

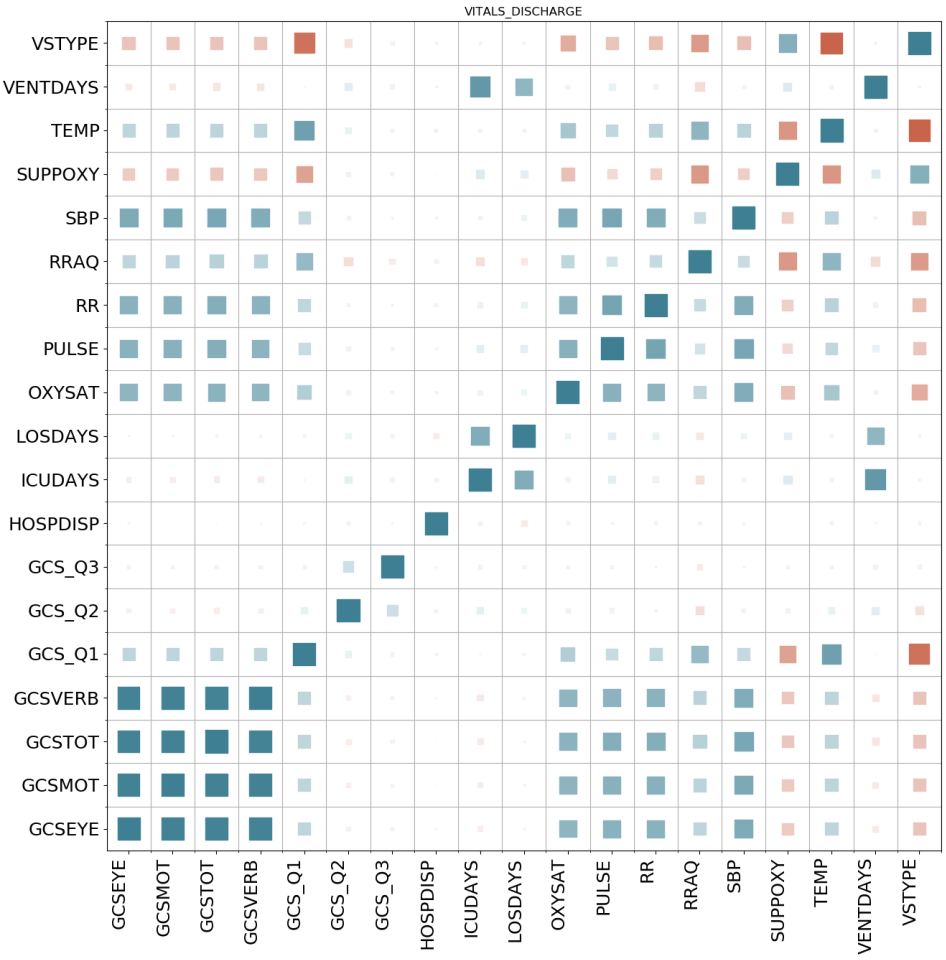


Figure 5.14.: VITALS and DISCHARGE correlation

To conclude, as no strong correlations between two independent variables were found, we were unable to define a model in order to predict those theoretical relations.

6. Conclusion

Finally, closing this Thesis, we are going to evaluate the work achieved, and ideas for future improvements or other possible lines of research will be provided.

In the first place, we have achieved analyzing the correlation of the tables included in the American trauma database TQP PUF of the year 2016, which has records of patients from 2006 to 2016.

Moreover, the thorough data analysis within the database provided with interesting remarks about some of most life threatening traumas. It also provided insight in the importance of patient vitals and some actuation protocols, such as the problems found when trying to obtain accurate trauma scores. Furthermore, we could see the differences in patients vital signs and their trauma scores between the different types of traumas.

Even so, the correlation analysis of the database proved to be unfruitful, as no significant correlation was found between variables not inherently related (for example, the birth year and the age). Nevertheless, the correlation analysis was helpful in confirming several claims that other studies had made about some situations in the hospital, such as the relation between days in the ICU and days in a ventilator, which we also found in the data analysis section (Section 4.1).

However, in addition to these general objectives of the Master Thesis, I also believe it is important to reflect that in this development process a parallel learning of areas and technologies which I had not previously worked with has been carried out, such as the handling of trauma in the hospitals, and the techniques involved in it, as well as a deeper understanding in data analysis methods.

6.1. Future developments

As the correlation analysis proved unfruitful, to finalize this Thesis I would like to highlight a series of ideas and improvements that could be carried out in order to improve results or extend functionalities that, due to lack of time, have not been included in this work.

- Obtain access to another trauma database so that the data can be crossed for analysis and evidence. In our case preferably the RETRAUCI trauma database.
- Expand the data analysis fraction to include more life threatening traumas, as well as dig deeper in those traumas, in search of abnormalities that may prove useful to improve the process of care.
- Use regression analysis and other machine learning techniques to try and find a suitable prediction model for the patient conditions while in the hospital.

Expanding the techniques to incorporate deep learning models would also be interesting.

- Incorporate the finding onto a medical simulation system and study its impact in teaching and improving professionals skills.

Bibliography

[1] F. Alberdi, I. García, L. Atutxa, and M. Zabarte. Epidemiology of severe trauma. *Medicina Intensiva (English Edition)*, 38(9):580–588, December 2014. ISSN 2173-5727. doi: 10.1016/j.medine.2014.06.002.

[2] *Trauma Quality Programs Participant Use File (PUF)*. . URL `https://www.facs.org/quality-programs/trauma/tqp/center-programs/ntdb/datasets`. Publication Title: American College of Surgeons.

[3] *WHO \textbar World Health Statistics 2019: Monitoring health for the SDGs*. . URL `http://www.who.int/gho/publications/world_health_statistics/2019/en/`. Publication Title: WHO.

[4] L C Harlan, W R Harlan, and P E Parsons. The economic impact of injuries: a major source of medical costs. *American Journal of Public Health*, 80(4):453–459, April 1990. ISSN 0090-0036.

[5] *WISQARS (Web-based Injury Statistics Query and Reporting System)\textbarInjury Center\textbarCDC*. March 2020. URL `https://www.cdc.gov/injury/wisqars/index.html`.

[6] *Defunciones por causas*. . URL `https://www.ine.es/jaxiT3/Tabla.htm?t=7947`. Publication Title: INE.

[7] Cdr Jamie L. Fitch, Paul T. Albini, Anish Y. Patel, Matthew S. Yanoff, Christian S. McEvoy, Chad T. Wilson, James Suliburk, Stephanie D. Gordy, and S. Rob Todd. Blunt versus penetrating trauma: Is there a resource intensity discrepancy? *American Journal of Surgery*, 218(6):1201–1205, 2019. ISSN 1879-1883. doi: 10.1016/j.amjsurg.2019.08.018.

[8] Nele Brusselaers, Stan Monstrey, Dirk Vogelaers, Eric Hoste, and Stijn Blot. Severe burn injury in europe: a systematic review of the incidence, etiology, morbidity, and mortality. *Critical Care*, 14(5):R188, October 2010. ISSN 1364-8535. doi: 10.1186/cc9300.

[9] *Traumatic Injury*. December 2012. URL `https://ufhealth.org/traumatic-injury`. Publication Title: UF Health, University of Florida Health.

[10] Justin Sobrino and Shahid Shafi. Timing and causes of death after injuries. *Proceedings (Baylor University. Medical Center)*, 26(2):120–123, April 2013. ISSN 0899-8280.

[11] John B. Kortbeek, Saud A. Al Turki, Jameel Ali, Jill A. Antoine, Bertil Bouillon, and Karen Brasel. Advanced trauma life support, 8th edition, the evidence for change. *The Journal of Trauma*, 64(6):1638–1650, June 2008. ISSN 1529-8809. doi: 10.1097/TA.0b013e3181744b03.

[12] Mark R. Hemmila. Management of the Injured Patient. In Gerard M. Doherty, editor, *Current Diagnosis & Treatment: Surgery*. McGraw Hill LLC, New York, NY, 15 edition, 2020.

[13] Pablo Perel, Ian Roberts, Omar Bouamra, Maralyn Woodford, Jane Mooney, and Fiona Lecky. Intracranial bleeding in patients with traumatic brain injury: A prognostic study. *BMC Emergency Medicine*, 9(1):15, August 2009. ISSN 1471-227X. doi: 10.1186/1471-227X-9-15.

[14] Thomedi Ventura, Cynthia Harrison-Felix, Nichole Carlson, Carolyn Diguiseppi, Barbara Gabella, Allen Brown, Michael Devivo, and Gale Whiteneck. Mortality after discharge from acute care hospitalization with traumatic brain injury: a population-based study. *Archives of Physical Medicine and Rehabilitation*, 91 (1):20–29, January 2010. ISSN 1532-821X. doi: 10.1016/j.apmr.2009.08.151.

[15] C. Clay Cothren, Ernest E. Moore, Holly B. Hedegaard, and Katy Meng. Epidemiology of urban trauma deaths: a comprehensive reassessment 10 years later. *World Journal of Surgery*, 31(7):1507–1511, July 2007. ISSN 0364-2313. doi: 10.1007/s00268-007-9087-2.

[16] Marc Maegele, Rolf Lefering, Nedim Yucel, Thorsten Tjardes, Dieter Rixen, Thomas Paffrath, Christian Simanski, Edmund Neugebauer, Bertil Bouillon, and AG Polytrauma of the German Trauma Society (DGU). Early coagulopathy in multiple injury: an analysis from the German Trauma Registry on 8724 patients. *Injury*, 38(3):298–304, March 2007. ISSN 0020-1383. doi: 10.1016/j.injury.2006.10.003.

[17] Ernest E. Moore, M. Margaret Knudson, Gregory J. Jurkovich, John J. Fildes, and J. Wayne Meredith. Emergency traumatologist or trauma and acute care surgeon: decision time. *Journal of the American College of Surgeons*, 209(3):394–395, September 2009. ISSN 1879-1190. doi: 10.1016/j.jamcollsurg.2009.06.003.

[18] Donald D. Trunkey and Robert C. Lim. Analysis of 425 consecutive trauma fatalities: An autopsy study. *Journal of the American College of Emergency Physicians*, 3(6):368–371, November 1974. ISSN 0361-1124. doi: 10.1016/S0361-1124(74)80005-5.

[19] T. A. Brennan, L. L. Leape, N. M. Laird, L. Hebert, A. R. Localio, A. G. Lawthers, J. P. Newhouse, P. C. Weiler, and H. H. Hiatt. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. *The New England Journal of Medicine*, 324(6):370–376, February 1991. ISSN 0028-4793. doi: 10.1056/NEJM199102073240604.

[20] Molla Sloane Donaldson. An Overview of To Err is Human: Re-emphasizing the Message of Patient Safety. In Ronda G. Hughes, editor, *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*, Advances in Patient Safety. Agency for Healthcare Research and Quality (US), Rockville (MD), 2008.

[21] *The Glasgow structured approach to assessment of the Glasgow Coma Scale.* . URL `https://www.glasgowcomascale.org/`.

[22] E. Wesley Ely, Brenda Truman, Ayumi Shintani, Jason W. W. Thomason, Arthur P. Wheeler, Sharon Gordon, Joseph Francis, Theodore Speroff, Shiva Gautam, Richard Margolin, Curtis N. Sessler, Robert S. Dittus, and Gordon R. Bernard. Monitoring Sedation Status Over Time in ICU Patients: Reliability and Validity of the Richmond Agitation-Sedation Scale (RASS). *JAMA*, 289 (22):2983–2991, June 2003. ISSN 0098-7484. doi: 10.1001/jama.289.22.2983.

[23] Anthony Marmarou, Juan Lu, Isabella Butcher, Gillian S. McHugh, Gordon D. Murray, Ewout W. Steyerberg, Nino A. Mushkudiani, Sung Choi, and Andrew I. R. Maas. Prognostic value of the Glasgow Coma Scale and pupil reactivity in traumatic brain injury assessed pre-hospital and on enrollment: an IMPACT analysis. *Journal of Neurotrauma*, 24(2):270–280, February 2007. ISSN 0897-7151. doi: 10.1089/neu.2006.0029.

[24] Efthimios J. Kouloulas, Alexandros G. Papadeas, Xanthi Michail, Damianos E. Sakas, and Efstathios J. Boviatsis. Prognostic value of time-related Glasgow coma scale components in severe traumatic brain injury: a prospective evaluation with respect to 1-year survival and functional outcome. *International Journal of Rehabilitation Research. Internationale Zeitschrift Fur Rehabilitationsforschung. Revue Internationale De Recherches De Readaptation*, 36(3):260–267, September 2013. ISSN 1473-5660. doi: 10.1097/MRR.0b013e32835fd99a.

[25] Rostam Jalali and Mansour Rezaei. A Comparison of the Glasgow Coma Scale Score with Full Outline of Unresponsiveness Scale to Predict Patients' Traumatic Brain Injury Outcomes in Intensive Care Units. *Critical Care Research and Practice*, 2014, 2014. ISSN 2090-1305. doi: 10.1155/2014/289803.

[26] Kishor Khanal, Sanjeeb Sudarshan Bhandari, Ninadini Shrestha, Subhash Prasad Acharya, and Moda Nath Marhatta. Comparison of outcome predictions by the Glasgow coma scale and the Full Outline of UnResponsiveness score in the neurological and neurosurgical patients in the Intensive Care Unit. *Indian Journal of Critical Care Medicine : Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine*, 20(8):473–476, August 2016. ISSN 0972-5229. doi: 10.4103/0972-5229.188199.

[27] *Abbreviated Injury Scale (AIS).* . URL `https://www.aaam.org/abbreviated-injury-scale-ais/`. Publication Title: Association for the Advancement of Automotive Medicine.

[28] Susan P. Baker, Brian O'neill, William Jr Haddon, and William B. Long. The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. *Journal of Trauma and Acute Care Surgery*, 14(3): 187–196, March 1974. ISSN 2163-0755.

[29] Wayne S. Copes, Howard R. Champion, William J. Sacco, Mary M. Lawnick, Susan L. Keast, and Lawrence W. Bain. The Injury Severity Score Revisited. *Journal of Trauma and Acute Care Surgery*, 28(1):69–77, January 1988. ISSN 2163-0755.

[30] Joshua B. Brown, Mark L. Gestring, Christine M. Leeper, Jason L. Sperry, Andrew B. Peitzman, Timothy R. Billiar, and Barbara A. Gaines. Characterizing injury severity in non-accidental trauma: Does injury severity score miss the mark? *The journal of trauma and acute care surgery*, 85(4):668–673, October 2018. ISSN 2163-0755. doi: 10.1097/TA.0000000000001841.

[31] André Lavoie, Lynne Moore, Natalie LeSage, Moishe Liberman, and John S. Sampalis. The New Injury Severity Score: A More Accurate Predictor of In-Hospital Mortality than the Injury Severity Score. *Journal of Trauma and Acute Care Surgery*, 56(6):1312–1320, June 2004. ISSN 2163-0755. doi: 10.1097/01.TA.0000075342.36072.EF.

[32] Qiangyu Deng, Bihan Tang, Chen Xue, Yuan Liu, Xu Liu, Yipeng Lv, and Lulu Zhang. Comparison of the Ability to Predict Mortality between the Injury Severity Score and the New Injury Severity Score: A Meta-Analysis. *International Journal of Environmental Research and Public Health*, 13(8):825, August 2016. doi: 10.3390/ijerph13080825.

[33] Howard Champion, William Sacco, Anthony Carnazzo, Wayne Copes, and William Fouty. Trauma score. *Critical Care Medicine*, 9(9):672–676, September 1981. ISSN 0090-3493.

[34] Howard R. Champion, William J. Sacco, Wayne S. Copes, Donald S. Gann, Thomas A. Gennarelli, and Maureen E. Flanagan. A Revision of the Trauma Score. *Journal of Trauma and Acute Care Surgery*, 29(5):623–629, May 1989. ISSN 2163-0755.

[35] B. Bouillon, E. Neugebauer, D. Rixen, R. Lefering, and T. Tiling. Value of clinical scoring systems for evaluation of injury severity and as an instrument for quality management of severely injured patients. *Zentralblatt Fur Chirurgie*, 121 (11):914–923, 1996. ISSN 0044-409X.

[36] Carl R. Boyd, Mary Ann Tolson, and Wayne S. Copes. Evaluating Trauma Care: The TRISS Method. *Journal of Trauma and Acute Care Surgery*, 27(4):370–378, April 1987. ISSN 2163-0755.

[37] Rameshbabu Homanna Javali, Krishnamoorthy, Akkamahadevi Patil, Madhu Srinivasarangan, Suraj, and Sriharsha. Comparison of Injury Severity Score, New Injury Severity Score, Revised Trauma Score and Trauma and Injury Severity Score for Mortality Prediction in Elderly Trauma Patients. *Indian Journal of Critical Care Medicine : Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine*, 23(2):73–77, February 2019. ISSN 0972-5229. doi: 10.5005/jp-journals-10071-23120.

[38] Jaspal Singh, Gulzar Gupta, Ramneesh Garg, and Ashish Gupta. Evaluation of trauma and prediction of outcome using TRISS method. *Journal of Emergencies, Trauma and Shock*, 4(4):446–449, 2011. ISSN 0974-2700. doi: 10.4103/0974-2700.86626.

[39] Joseph Lee Rodgers and W Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[40] G.U. Yule and M.G. Kendall. *An introduction to the theory of statistics*. Hafner Pub. Co., 1950.

[41] Babak Mahdavi-Damghani. The Non-Misleading Value of Inferred Correlation: An Introduction to the Cointelation Model. SSRN Scholarly Paper ID 2429120, Social Science Research Network, Rochester, NY, 2013.

[42] D. G. Altman and J. M. Bland. Measurement in Medicine: The Analysis of Method Comparison Studies. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(3):307–317, 1983. ISSN 1467-9884. doi: 10.2307/2987937.

[43] Man Hung, Jerry Bounsanga, and Maren Wright Voss. Interpretation of correlations in clinical research. *Postgraduate medicine*, 129(8):902–906, November 2017. ISSN 0032-5481. doi: 10.1080/00325481.2017.1383820.

[44] Walid Kamal Abdelbasset, Saud F. Alsubaie, Sayed A. Tantawy, Tamer I. Abo Elyazed, and Ahmed A. Elshehawy. A cross-sectional study on the correlation between physical activity levels and health-related quality of life in community-dwelling middle-aged and older adults. *Medicine*, 98(11):e14895, March 2019. ISSN 0025-7974. doi: 10.1097/MD.0000000000014895.

[45] Hélio Amante Miot and Hélio Amante Miot. Correlation analysis in clinical and experimental studies. *Jornal Vascular Brasileiro*, 17(4):275–279, December 2018. ISSN 1677-5449. doi: 10.1590/1677-5449.174118.

[46] J.M. Steele. *The Cauchy-Schwarz Master Class ICM Edition: An Introduction to the Art of Mathematical Inequalities*. Cambridge University Press, 2010. ISBN 978-0-521-17001-7.

[47] Ann Lehman, Norm O'Rourke, Larry Hatcher, and Edward Stepanski. *JMP for Basic Univariate and Multivariate Statistics: A Step-by-Step Guide*. February 2005. ISBN 1-59047-576-3.

[48] Morgan Shields. *Research Methodology and Statistical Methods*. Scientific e-Resources, August 2019. ISBN 978-1-83947-332-6.

[49] Helena Chmura Kraemer. Correlation coefficients in medical research: from product moment correlation to the odds ratio. *Statistical Methods in Medical Research*, 15(6):525–545, December 2006. ISSN 0962-2802. doi: 10.1177/0962280206070650.

[50] Frank E. Grubbs. Procedures for Detecting Outlying Observations in Samples. *Technometrics*, 11(1):1–21, February 1969. ISSN 0040-1706. doi: 10.1080/00401706.1969.10490657.

[51] Brian D. Ripley. Robust Statistics. 2004. doi: 10.4135/9781412953948.n402.

[52] Chandan Mukherjee, Howard White, and Marc Wuyts. *Econometrics and Data Analysis for Developing Countries*. Routledge, September 2013. ISBN 978-1-136-14460-8.

[53] Arthur Zimek and Peter Filzmoser. There and back again: Outlier detection between statistical reasoning and data mining algorithms. *WIREs Data Mining and Knowledge Discovery*, 8(6):e1280, 2018. ISSN 1942-4795. doi: 10.1002/widm.1280.

[54] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):363–387, 2012. ISSN 1932-1872. doi: 10.1002/sam.11161.

[55] Victoria Hodge and Jim Austin. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2):85–126, October 2004. ISSN 1573-7462. doi: 10.1023/B:AIRE.0000045502.10941.a9.

[56] Cátia M. Salgado, Carlos Azevedo, Hugo Proença, and Susana M. Vieira. Noise Versus Outliers. In *Secondary Analysis of Electronic Health Records*. Springer, Cham (CH), 2016. ISBN 978-3-319-43740-8 978-3-319-43742-2.

[57] Helge Erik Solberg and Ari Lahti. Detection of outliers in reference distributions: performance of Horn's algorithm. *Clinical chemistry*, 51(12):2326–2332, 2005.

[58] H. Beyer. Tukey, John W.: Exploratory Data Analysis. Addison-Wesley Publishing Company Reading, Mass. — Menlo Park, Cal., London, Amsterdam, Don Mills, Ontario, Sydney 1977, XVI, 688 S. *Biometrical Journal*, 23(4):413–414, 1981. ISSN 1521-4036. doi: 10.1002/bimj.4710230408.

[59] Edwin Knorr, Raymond Ng, and Vladimir Tucakov. Distance-Based Outliers: Algorithms and Applications. *The VLDB Journal*, 8:237–253, February 2000. doi: 10.1007/s007780050006.

[60] Markus Breunig, Hans-Peter Kriegel, Raymond Ng, and Joerg Sander. *LOF: Identifying Density-Based Local Outliers.*, volume 29. June 2000. doi: 10.1145/342009.335388.

[61] Arthur Zimek and Erich Schubert. Outlier Detection. *Encyclopedia of Database Systems*, 2017. doi: 10.1007/978-1-4899-7993-3_80719-1.

[62] E. Kreyszig. *Advanced Engineering Mathematics*. Wiley, 1999. ISBN 978-0-471-15496-9.

[63] Christopher P. Carroll, Joseph A. Cochran, Janet P. Price, Clare E. Guse, and Marjorie C. Wang. The AIS-2005 Revision in Severe Traumatic Brain Injury: Mission Accomplished or Problems for Future Research? *Annals of Advances in Automotive Medicine / Annual Scientific Conference*, 54:233–238, January 2010. ISSN 1943-2461.

[64] *ICD - ICD-10-CM - International Classification of Diseases,(ICD-10-CM/PCS Transition)*. March 2019. URL `https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm`.

[65] *Python*. . URL `https://www.python.org/`. Publication Title: Python.org.

[66] *SciPy.org*. . URL `https://www.scipy.org/`.

[67] *pandas - Python Data Analysis Library*. . URL `https://pandas.pydata.org/`.

[68] *SQLite*. . URL `https://www.sqlite.org/index.html`.

[69] *Most Widely Deployed SQL Database Engine. SQLite.org.* . URL `https://sqlite.org/mostdeployed.html`. Publication Title: https://sqlite.org/.

[70] Roland N. Pittman. *Oxygen Transport in Normal and Pathological Situations: Defects and Compensations*. Morgan & Claypool Life Sciences, 2011.

[71] Jesús Villar, Carlos Ferrando, and Robert M Kacmarek. Managing Persistent Hypoxemia: what is new? *F1000Research*, 6, November 2017. ISSN 2046-1402. doi: 10.12688/f1000research.11760.1.

[72] Jesse J Corry. Use of hypothermia in the intensive care unit. *World Journal of Critical Care Medicine*, 1(4):106–122, August 2012. ISSN 2220-3141. doi: 10.5492/wjccm.v1.i4.106.

[73] Hannah Wunsch, Jason Wagner, Maximilian Herlim, David Chong, Andrew Kramer, and Scott D. Halpern. ICU Occupancy and mechanical ventilator use in the United States. *Critical care medicine*, 41(12), December 2013. ISSN 0090-3493. doi: 10.1097/CCM.0b013e318298a139.

[74] Michel Heijnen and Shirley Rietdyk. Falls in young adults: Perceived causes and environmental factors assessed with a daily online survey. *Human Movement Science*, 46:86–95, April 2016. doi: 10.1016/j.humov.2015.12.007.

[75] R. P. Heaney, J. C. Gallagher, C. C. Johnston, R. Neer, A. M. Parfitt, and G. D. Whedon. Calcium nutrition and bone health in the elderly. *The American Journal of Clinical Nutrition*, 36(5):986–1013, November 1982. ISSN 0002-9165. doi: 10.1093/ajcn/36.5.986.

[76] José Gustavo Parreira, André Mazzini Ferreira Vianna, Gabriel Silva Cardoso, Walter Zavem Karakhanian, Daniela Calil, Jaqueline A. Giannini Perlingeiro, Silvia C. Soldá, and José Cesar Assef. Severe injuries from falls on the same level. *Revista Da Associacao Medica Brasileira (1992)*, 56(6):660–664, December 2010. ISSN 1806-9282. doi: 10.1590/s0104-42302010000600013.

[77] D. W. Bates, N. Spell, D. J. Cullen, E. Burdick, N. Laird, L. A. Petersen, S. D. Small, B. J. Sweitzer, and L. L. Leape. The costs of adverse drug events in hospitalized patients. Adverse Drug Events Prevention Study Group. *JAMA*, 277(4):307–311, January 1997. ISSN 0098-7484.

[78] Philip H. Pucher, Rajesh Aggarwal, Ahmed Twaij, Nicola Batrick, Michael Jenkins, and Ara Darzi. Identifying and Addressing Preventable Process Errors in Trauma Care. *World Journal of Surgery*, 37(4):752–758, April 2013. ISSN 1432-2323. doi: 10.1007/s00268-013-1917-9.

A. Ethical, economic, social, and environmental aspects

A.1. Introduction

Data analysis is being used in more and more fields each day, as it has proven that upgrading existing protocols and services is possible with its help. In the medical field it has opened the door to machine learning, and in the most broad sense artificial intelligence. This Master Thesis focused on analyzing the data stored in a database of trauma patients records.

A.2. Description of relevant project related problems

- **Social impact:** This Master Thesis has a direct social impact, as the findings can help Emergency Medical Services (EMS) and Emergency Department (ED) departments to be better prepared to face problems, and to respond quicker to them. Therefore, increasing the patients survival probability.
- **Economical impact:** The economical impact of this Master Thesis is included within the costs to the hospital from trauma related injuries. A study [77], found that the annual cost of all adverse drug events and preventable adverse drug events, in a 700 bed teaching hospital, was estimated to be 5,6 million dollars and 2,8 million dollars respectively.
- **Ethical impact:** Pucher et al. [78] claims that medical errors cause injuries to healthy patients in approximately 3,6% of hospital patients, with a risk of harm of 30%. Moreover, the majority of the errors (51%) occurred during the initial stay at the ED. With better clinical training these errors could be significantly reduced.
- **Environmental impact:** There is no environmental impact in the realization of this Master Thesis, as it was all done within a computer environment.

A.3. Conclusions

The improvement of the clinical simulation thanks to data analysis would, in turn, improve the care that patients receive during their episode of trauma, as the EMS and ED services would be better prepared. Furthermore, it would also save the hospital money from misuse of drugs, and possible lawsuits from the patients or families due to medical negligence.

B. Economic budget

In Table B.1 the economic budget is explained.

Workforce cost (Direct cost)			
	Hours	**Price/hour**	**Total**
	400	12	4.800 €
Material costs (Direct cost)			
	Price	**Use in months**	**Amortization in years**	**Total**
Personal computer (Software included)	1.000,00 €	4	5	66,66 €
TQP PUF AY 2016 trauma database	500,00 €	4	5	33,33 €
Total of material costs				99,99 €
Total costs			
Workforce costs				4.800 €
Material costs				99,99 €
General costs (15% of direct costs) (Indirect cost)				734,99 €
Industrial benefit (6% of direct and indirect costs)				338,09 €
IVA (21%)				1254,34 €
Total				7.227,41 €

Table B.1.: Economic budget of the Thesis