NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF MATHEMATICS

FRANCISCO BARROSO DE SOUSA ALBARDEIRO

Bachelor in Mathematics

CREATING A LUNG ULTRASOUND INDEX TO PREDICT COVID-19 SEVERITY

A CONTRIBUTION TO AID MEDICAL PERSONNEL DECISION MAKING

MASTER IN MATHEMATICS AND APPLICATIONS

NOVA University Lisbon
September, 2022

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF MATHEMATICS

CREATING A LUNG ULTRASOUND INDEX TO PREDICT COVID-19 SEVERITY

A CONTRIBUTION TO AID MEDICAL PERSONNEL DECISION MAKING

FRANCISCO BARROSO DE SOUSA ALBARDEIRO

Bachelor in Mathematics

Adviser: Regina Maria Baltazar Bispo
Assistant Professor, NOVA School of Science and Technology, Universidade NOVA de Lisboa

Co-adviser: Maria Isabel Azevedo Rodrigues Gomes
Associated Professor, NOVA School of Science and Technology, Universidade NOVA de Lisboa

Examination Committee

Chair: Name of the committee chairperson
Full Professor, FCT-NOVA

Rapporteur: Name of a rapporteur
Associate Professor, Another University

Members: Another member of the committee
Full Professor, Another University

Yet another member of the committee
Assistant Professor, Another University

MASTER IN MATHEMATICS AND APPLICATIONS
SPECIALIZATION IN STATISTICS AND OPERATIONS RESEARCH

NOVA University Lisbon
September, 2022

Creating a Lung Ultrasound Index To Predict Covid-19 Severity

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v6.9.4) [**novathesis-manual**].

To my Avó Gé, Avó Bia and Avô Luís,

Acknowledgements

First, i would like to thank both my supervisors, Prof. Regina Bispo and Prof Isabel Gomes. Their demand and support were crucial to the development of this work, and their human touch was fundamental to my growth as a person, for which i will always be thankful. I would also like to thank the Center for Mathematics and Application, CMA, for the financial support given, in the form of a research scholarship, while developing this thesis. To the FCT-UNL, in specific to the Department of Mathematics, i would like to express my gratitude for all the knowledge and for the great studying environment.
On a personal level, first i would like to thank my working friend, Filipe Sousa, for all the hours of motivation and companionship. To my girlfriend, Sofiya, and my family, i thank for the patience to listen to my frustrations and problems and for providing me the best home conditions to develop this work. Finally, to everyone that, in one way or another, helped me in my academic course, a big thank you to all.

"Do or do not, there is no try." (Yoda, Star Wars)

Abstract

The COVID-19 pandemic led to numerous challenges related to the management of resources in the health infrastructures. One particular case was the congestion of ICU in hospitals and the small amount of available ventilators. The decisions of the medical personnel were of extreme responsibility and, therefore, the usage and understanding of all the available information was crucial for these workers.
The study objects of this work are lung ultrasounds, an exam used in hospitals to assess the state of the lung, that are useful when patients develop pneumonia, which is a possible consequence of COVID-19 infection. Lung ultrasounds provide information about B-lines, about pleura homogeneity and about possible consolidations. The aim of this study is to create an index, based on this collected information, that can predict death and the need for ICU hospitalization for COVID-19 hospitalized patients. The index is then compared to the Lung Ultrasound Score (LUS), already in use.
To address this challenge a Severity Index (SI) is developed, based on the patients' hospital length of stay, ICU hospitalization and outcome (death or discharge). This index is then used as the response variable in three Generalized Linear Models. The first assumes that the SI follows a Gamma distribution and uses the inverse link function (G-Inv), the second also assumes the index follows a Gamma distribution but uses the log link function (G-Log) and the last assumes that the SI follows a Log-normal distribution and therefore the logarithm of the SI is taken, originating a Normal distribution, and the identity link function is used (LN). The models are then compared among themselves and with LUS, the score used currently in the hospital. The data used is from COVID-19 patients hospitalized in Garcia de Orta hospital, in Almada.
The results point that the developed models clearly outperform LUS in death prediction, with AUC's (obtained after the construction of the ROC curves) of 0.756, 0.768 and 0.768 for G-Inv, G-Log and LN, respectively, compared to the 0.641 for LUS. Regarding ICU prediction, LUS has the best performance, with AUC of 0.875, followed by LN (0.82), G-Log (0.818) and G-Inv (0.752). After calculating the ideal thresholds for each model, for both death and ICU prediction, the most balanced model is G-Log. The ICU threshold is 21.2, with sensitivity of 80.6% and specificity of 72.9%, and the death threshold is 32.9,

with sensitivity of 68.8% and specificity of 86%.

Keywords: COVID-19, Statistics, Generalized Linear Models, Lung Ultrasound, Index Development

Resumo

A pandemia de COVID-19 originou vários desafios relacionados com a gestão de recursos em infraestruturas de saúde. Um caso particular foi a congestão existente em UCI nos hospitais e a pequena disponibilidade de ventiladores. As decisões tomadas pelo pessoal médico são de extrema responsabilidade e, por isso, o uso e compreensão de toda a informação disponível é crucial para estes profissionais.

Os objetos de estudo deste trabalho são as ecografias pulmonares, um exame realizado nos hospitais para aferir o estado dos pulmões, que são muito úteis quando os pacientes desenvolvem pneumonia, uma possível consequência da infeção por COVID-19. As ecografias pulmonares disponibilizam informação acerca das linhas-B, da homogeneidade da pleura e sobre a possível existência de consolidações.

O âmbito deste estudo é a criação de um indicador, baseado na informação recolhida, que possa prever a morte e a necessidade de internamento em UCI para pacientes internados com COVID-19. O indicador será no final comparado com o Lung Ultrasound Score (LUS), Pontuação da Ecografia Pulmonar, que é utilizado atualmente.

Para enfrentar este desafio, foi construído um indicador de gravidade (SI), baseado no tempo de internamento, internamento em UCI e no desfecho do internamento (morte ou alta) dos pacientes estudados. Este indicador é depois usado como variável de resposta em três Modelos Lineares Generalizados. O primeiro assume que o indicador segue uma distribuição Gamma, e usa a função de ligação inversa (G-inv), o segundo também assume uma distribuição Gamma mas usa a função de ligação logarítmica (G-Log) e o último assume que o SI segue uma distribuição Log-Normal e por isso, usa o logarítmo do SI, originando uma distribuição Normal, e utiliza a função de ligação identidade (LN). Os modelos são comparados entre eles e também com o LUS. A informação utilizada provém de pacientes hospitalizados com COVID-19 no hospital García de Orta, em Almada.

Os resultados mostram que os modelos construídos são superiores ao LUS na previsão de morte, com AUC's (obtidos depois da construção de curvas ROC) de 0.756 (G-Inv), 0.768 (G-Log) e 0.768 (LN), comparado com 0.641 para o LUS. Relativamente à previsão de internamento em UCI, LUS tem o melhor desempenho, com um AUC de 0.875, seguido do LN (0.82), G-Log (0.818) e G-Inv (0.752). Depois de calculados os limiares ideais para

cada modelo, quer para a previsão de morte quer de internamento em UCI, o modelo mais equilibrado é o G-Log. O limiar para UCI é 21.2, gerando uma sensibilidade de 80.6% e uma especificidade de 72.9%. O limiar para a morte é 32.9, com uma sensibilidade de 68.8% e uma especificidade de 86%.

Palavras-chave: COVID-19, Estatística, Modelos Lineares Generalizados, Ecografias Pulmonares, Desenvolvimento de Indicadores

Contents

List of Figures

List of Tables

Acronyms

AIC Akaike Information Criterion 27–29
AUC Area Under the Curve 5, 7, 8, 10, 11, 36, 37, 57, 61, 78, 79, 82

CT Computerized Tomography 8, 10
Ct Cycle threshold 3, 4

G-Inv Generalized Linear Model assuming the Scores follow a Gamma distribution with inverse link function 68, 78–82
G-Log Generalized Linear Model assuming the Scores follow a Gamma distribution with log link function 68, 78–82
GLM Generalized Linear Model 2, 5, 6, 12, 17, 21, 24, 30, 41, 53, 68, 75, 81
GOF Goodness of Fit 2, 26, 27

HOMALS Homogeneity Analysis by Means of Alternating Least Squares 34

ICU Intensive Care Unit 1, 2, 5, 8, 10, 11, 13, 41–44, 50, 52–54, 56, 57, 59–61, 78–82

KS Kolmogorov-Smirnov 15, 16, 41, 55, 56, 76

LN Generalized Linear Model assuming the logarithm of the Scores follow Normal distribution with identity link function 68, 78–82
Lob Lobar Consolidations 14, 38, 53, 64–67
LOS Length of Stay 1, 2, 5, 41–44, 47, 49, 50, 52–56, 81, 82
LUS Lung Ultrasound Score 2, 10, 12, 78–82

NPS Nasopharyngeal Swabs 3, 4

OPS Oropharyngeal Swabs 4

PC Principal Component 32, 65–67, 69, 73, 76

PCA Principal Components Analysis 2, 31, 33, 34, 63, 65

PRINCALS Principal Components Analysis by means of Alternating Least Squares 35, 63, 65

ROC Receiver Operating Characteristic 2, 5, 8, 36, 57, 78, 82

SI Severity Index 54

SubP Subpleural Consolidations 14, 38, 53, 63, 66

TD Time until Death 55, 56, 59

TUD Time Until Discharge 55, 56

TUID Time Until ICU Discharge 55–57, 61

1 Introduction

Healthcare is one of the most important areas of the current society, something that was made clear by the COVID-19 pandemic. One of the main jobs of medical staff is to heal patients. However, in order to be able to do so it is of extreme importance that a good diagnostic is made, so the best treatment can be executed. In the current world of ever evolving technology along with ease to obtain data, there are a lot of little explored combinations between the information obtained by the diagnostic procedures and mathematical tools to take full advantage of them. One of the fields where statistical techniques can be very useful is in predicting the likelihood of certain outcomes given some patient information like blood test results, symptom description or, as is the case in this thesis, ultrasound scores. This can be achieved by deriving an index based on a sample of patients where the outcome is known. The outcomes are related to the patient's Length of Stay (LOS), whether there was the need for Intensive Care Unit (ICU) hospitalization and for how long and finally, whether the patient died or recovered. The ongoing pandemic made clear that risk stratification is central to the functionality and organization of hospital units, therefore the effectiveness of the predictions of these indexes can be of great aid for decision makers, especially in situations where the hospital's patients capacity is being put to the test, like it was in the critical phases of the pandemic.

1.1 Objectives

The creation of a lung ultrasound index for predicting Covid-19 severity is the central focus of this study. The goal is that this index can have a high correlation with death, hospital LOS and ICU hospitalization so that it can be a helpful tool in managing the hospital's professional and material resources. There are also other objectives that will help the fulfillment of the primary objective. The first secondary objective is the development of a severity indicator, associated to each patients' lung ultrasounds, that will work as the response variable of the regression models. The goal is that this severity indicator integrates the known information regarding the final outcome of the patients, death or

recovering, but also about the LOS and whether there was the need for ICU hospitalization. The idea is that this severity indicator can be accurate regarding the patients' status when the ultrasounds were made, so the regression, made solely on the lung ultrasounds information, can have good correlation with the referred information. Other secondary objective is to understand, from the variables collected in the ultrasounds, which are more significant to assess Covid-19 severity. In practice, this will be done when the regression model is developed using stepwise algorithms and nested models comparisons. The final purpose of the study is to compare the regression results with Lung Ultrasound Score (LUS), the current scoring technique used by doctors in Hospital Garcia de Orta, to understand if the constructed models can be helpful in aiding medical decisions.

1.2 Dataset

The dataset has information about 141 lung ultrasounds, corresponding to 51 patients, where the number of ultrasounds per patient changes depending on the length of stay of each one. Each ultrasound has information about both lungs, with each lung separated in 4 different zones. Every zone has 4 variables described: B-lines, scored with 0 (normal), 1,2 or 3 (worst); Pleura, 0 if normal or 1 if heterogeneous; Subpleural Consolidations, 0 if absent or 1 if present; and Lobar consolidations, 0 if absent or 1 if present. Regarding each patient, besides sociodemographic information, there is data about its outcome (death or recovering), the length of stay, whether there was the need for ICU hospitalization and the respective ICU length of stay and about comorbidities. Finally, each lung ultrasound is scored according to LUS, the currently used index to assess lung ultrasound severity.

1.3 Structure

The thesis will be organized in 5 chapters, including this introductory one. Chapter 2 is the Literature Review where, at first, are exposed different methods, with its own advantages and disadvantages, used to diagnose COVID-19. After that, some disease severity models are explored, mainly based on Generalized Linear Model (GLM)'s, including survival analysis and logistic regressions. This is followed by the exploring of composite indicators built with the intention of evaluating the severity of different health problems. Finally, some studies aiming to predict COVID-19 severity are assessed as well as how medical imaging tools can help to make these predictions. The dataset and methodology are described in Chapter 3, with the main focus being on GLM's and their particularities. Principal Components Analysis (PCA), Goodness of Fit (GOF) tests and Receiver Operating Characteristic (ROC) curves definitions are also described. In Chapter 4 the results and their discussion are presented, it is explained how the proxy variable for disease severity is constructed and the application of GLM's is thoroughly described. In the end, the prediction quality is assessed as well as the comparison with the original score. Finally, Chapter 5 has the conclusions of this study.

2

Literature Review

The COVID-19 outbreak was declared as a pandemic on March 11^{th} 2020 by the World Health Organization, (WHO), and has had an enormous impact in many areas of our society throughout the world. This disease may manifest itself differently from host to host; some are asymptomatic, others have mild symptoms but there are also cases of severe problems that can lead to death.
This section summarizes a review of recent severity assessment techniques. First some COVID-19 diagnosis techniques, used to differentiate COVID and non-COVID patients. Then, are described some methods used in the medical area to predict the severity and possible outcomes in different diseases' patients. After that, COVID severity predicting tools are described, mainly those using mathematical approaches, and in the end it is described how medical imaging techniques can help to determine the probability of death or other unfavourable outcomes.

2.1 COVID Diagnosis Techniques

COVID-19 is a disease caused by a coronavirus and, in severe cases, can lead to pneumonia, a lung infection that affects mainly the alveoli [1]. Therefore, according to Miller et al. [2] the best procedure for laboratorial analysis is retracting specimens through nasal aspirates, nasal washes or Nasopharyngeal Swabs (NPS). After that a nucleic acid amplification test (NAAT) is performed on the sample, where Reverse Transcription Polymerase Chain Reaction (RT-PCR) is a common method for amplification [3]. These tests have a metric called Cycle threshold (Ct) that is *"the number of cycles required to amplify viral RNA to a detectable level, which provides a quantifiable estimate of viral load"*[4]. Less cycles lead to higher viral load. However, it is not always possible to assemble these optimal conditions when testing. Therefore, other methods for COVID-19 diagnosis were studied.

The beginning of the COVID-19 pandemic created a shortage of swabs capable of NPS and for that reason other possible zones to retract specimens were investigated. Lee et al. [5] made a meta-analysis using studies that compared NPS with other medical test

methods. Saliva specimens tests (ST) had a 88% positive prediction in true positive cases, compared to 94% in NPS and 79% of true positives were predicted correctly by both (dual prediction). This study also shows these proportions in symptomatic cases, ST predict correctly 88%, NPS 96% and the dual 82%; and for asymptomatic cases, 87%, 73% and 57%, respectively.
Regarding Oropharyngeal Swabs (OPS) the study showed similar proportions of correct positive predictions, 84% in OPS and 88% in NPS, but a relatively small agreement proportion, 68%. However, it must be noted that the NPS percentage is highly penalized in one of the studies where only 41% of the true positives are correctly predicted.
In respect of Nasal Swabs (NS), either mid-turbinate (MT) or anterior nares (AN), the prediction are considerably lower than when one uses NPS. NPS percentage positive is 98% compared to 82% of the NS but with a good agreement between them as dual percentage is 79%. Again in symptomatic cases the proportions are higher, but only slightly, 99%, 82% and 81% respectively.
Finally studies where a combination of OP and NS was used were the best performing ones. The positive percentage was equal to the NPS, 97%, and the dual positive percentage was 90%.

The type of tests referred in the previous paragraph also have some shortcomings: the need for expert personnel, the cost associated and the time frame between the test taking and the results [6]. Consequently, Bruzzone et al. [6] compared the sensitivity and specificity of antigen tests to the standard RT-PCR tests. The results showed that these have a high specificity, 100%, as no false positives were recorded and the sensitivity was 78.7%. However the sensitivity value change significantly depending on the Ct value associated to each test. Low value of Ct indicate higher viral load, and therefore when the Ct values are low (< 25) the sensitivity rises to 96.2%, which brings to debate which cut-off value should be used for a test to be positive. In this study 35 was the cut-off value used but the optimal value to maximize sensitivity was 29 in this case.

As these COVID tests require a new form of organization and logistics, some studies tried to use quicker and less costly tests to diagnose COVID. Abayomi-Alli et al. [7] used machine learning techniques applied to blood samples, a common hospital exam, to identify COVID cases. The best classifiers were Adaboost, ExtraTrees, Decision Tree, QDA, and random forest models achieving accuracies of 99.28%, and 99.28%, 98.5%, 94.6%, and 92.9%, respectively, accomplishing better results when compared to other studies.

Although the initial Covid-19 detection methods are very important for public health decisions, they are not capable of predicting if a patient will be asymptomatic or will need hospitalization, with the possible exception of the Ct value [8], and so other techniques are needed for this purpose.

2.2 Disease severity models

A very common way of evaluating the severity of a given disease is by using the hospital Length of Stay (LOS) as a proxy variable, in the sense that it has a very high correlation with the actual variable one wants to study. it is based on the assumption that the longer someone stays in the hospital the higher the severity of the disease.

Generalized Linear Model (GLM) have been used to help medical decision making based on predictions related to unfavourable outcomes. Survival analysis is a particular case of a GLM and is a common mathematical approach to study the LOS and the effect that some variables may, or may not have on it. Kwoh et al. [9] researched the effect that age, nationality, race, admission status and socio-economic conditions (measured by a patient subsidy status) had on LOS of stroke patients in Singapore by using Cox proportional hazard model, which is a semi-parametric model. The results showed that age, race and subsidy status had a significant influence on LOS, with LOS increasing with age while people that have subsidies also stay longer. Al Mamlook et al. [10] instead of using LOS as the response variable used time until death in patients with Lung Cancer. The goal was to investigate if Karnofsky Performance Status (KPS) and the Eastern Cooperative Oncology Group Performance Status Scale (ECOG PS) were good indicators of survival. KPS is scored from 0 to 100 and evaluates a patient's functional status. The results showed that this was not a good predictor, but it must be noted that there were no patients with scores between 0 and 25, and just a small fraction between 26 and 50. ECOG PS, on the other hand, showed promising results but had similar problems. This indicator is scored from 0 to 5, with 5 being the worst, and there was only 1 patient with score of 3 and none above or equal to 4.
Although survival analysis is a good tool to understand the time range associated with some diseases it has some challenges in terms of individual analysis and the type of care needed until hospital discharge or death. For example, [9] et al. doesn't differentiate patients that spent time in Intensive Care Unit (ICU) from those that didn't. It also doesn't account for the patient's status at the time of discharge (completely healthy or in need of home care).
Other GLM's may be used to create indexes or to make predictions related to health. Yamamoto et al. [11] used multivariate logistic regression, useful when the dependent variable is binary, to create the modifed Abbreviated Burn Severity Index (mABSI) a score based system with the goal of making predictions of in-hospital mortality for patients with inalition injuries. The study assessed that the significant predictive variables were age, self-inflicted injury, total cutaneous burned area (%) and if mechanical ventilation was needed. To validate the model, Receiver Operating Characteristic (ROC) curves were constructed on the validation cohort with Area Under the Curve (AUC) value of 0.94.
Stuhler et al. [12] applied Bayesian GLM to predict the outcome of patients when exposed to different treatment options for multiple sclerosis (MS) and from there choose the best

for each individual. Two response variables were used, first the number of on-therapy relapses modelled by a negative binomial distribution and then the confirmed disability progression (CDP) modelled as a binomial variable. The predictive variables were age, gender, EDSS (a metric used to quantify MS patient's incapacity), index therapy, current therapy, diagnosis distance, relapse distance, relapses count, number of therapies taken before the start of index therapy, whether a second-line therapy was taken before the start of index therapy, index duration and clinical site. The results showed that patients treated with the therapy that the model predicted to lead to the best outcome were less likely to suffer from relapses or to have CDP.
GLM's are an adequate methodology to model disease severity associated to different health issues that can work with response variables that follow distributions belonging to the exponential family, like binary variables or some skewed distributions.

2.3 Composite Indicators

Another technique to assess a person's health are composite indicators. Composite indicators work by analyzing several variables that may be related to the intended effect and translate them all into a single value that is compared to the associated scale.
Schoufour et al. [13] used this method to construct a Frailty Index (FI) related to the population over the age of 45, where frailty is *"defined as a state of vulnerability to adverse health outcomes at old age"*. The variables used to construct the index were related to some sort of deficit and the authors chose the ones that simultaneously verified three conditions: the deficit was health related; in general increased with age and was not too rare ($< 5\%$) nor too common ($> 80\%$). In general the variables took the value 1 if the deficit was present and 0 otherwise. Some of them, however, had the possibility of intermediate values, for example to separate mild and severe deficits (1 for severe deficit, 0.5 for mild deficit and 0 for no deficit). In the end the sum of all scores divided by the number of studied variables (45) originated the final score. The study showed that the Frailty Index was strongly associated with all cause mortality where the adjusted hazard ratio on mortality increased 1.05 per FI unit increase.
Bernabé et al. [14] manipulated a composite indicator related to dental health, T-health index, to assess which weights were more appropriate for the index to be a good indicator of oral health, using the proxy variable of perceived oral health. The possible categories for each tooth were "missing", "decayed", "filled" and "sound"(i.e., healthy). "Missing" had a weight of 0 and sound of 1, the "decayed" and "filled" weights were the ones changed but always respecting the hierarchy of a decayed tooth being worse (or equal) to a filled one. The results showed that best association between the T-health Index and the perceived oral health were achieved when the weight for a "decayed" tooth was half of the one for a "filled" tooth with the weight for a filled tooth not exceeding 0.5.
Composite indicators are a good tool to quantify some health related concepts because they are easily constructed and have a very straightforward interpretation. However, this

indexes have some downsides, the fact that sometimes there are no weights for each variable results in all variables having the same influence in the indicator, which, in most cases, isn't adequate. Even when weights are used, they are defined by subjective decisions which can lead to a biased indicator.

2.4 Predicting COVID severity

Since the early days of this pandemic that scientists started searching for the best ways to predict the consequences of the disease in patients [15].

Bolourani et al. [16] used three machine learning models that predicted respiratory failure within 48 hours of patient admission. The first model was based on the XGBoost, an algorithm that applies gradient boost method and decision trees; the second combined XGBoost with SMOTEENN, a method for oversampling using synthetic minority oversampling technique (SMOTE) and undersampling with edited nearest neighbors; and the last one was logistic regression. All three were compared between them and compared to a Modified Early Warning Score, a score built from heart rate, respiratory rate, systolic blood pressure and body temperature. The best performing model was XGBoost, with an AUC of 0.77 and some of the important variables used were *"type of oxygen delivery used in the emergency department, patient age, Emergency Severity Index level, respiratory rate, serum lactate, and demographic characteristics"*[16].

Chen et al. [17] built a risk score aiming to predict death related to COVID-19. First 64 variables were selected associated to each patient, then, after dealing with missing values, a random forest algorithm was used to select the main features, 21 variables were selected. The 21 variables were dichotomized for clinic utility and finally a Least Absolute Shrinkage and Selection Operator (LASSO) regression was applied to reach the final 8 risk factors. To develop the final risk score a Cox hazard regression analysis was performed where the significant factors were Oxygen saturation < 90%, blood Urea nitrogen > upper limit of normal (ULN), Respiratory rate > 30 (per minute), admission before the date of the first national Maximum number of daily new cases was reached, Age ≥ 60 years, Procalcitonin > ULN, C-reactive protein (CRP) > ULN and absolute Neutrophil counts > ULN. The AUC was of 0.90 when the score was applied to the testing cohort.

Das et al. [18] also used machine learning techniques to predict COVID community mortality risk. Five methods were tested based on four individual characteristics: age group, sex, province and exposure. The first three are all self-explanatory and exposure was considered based on if people went to hospitals, nursing homes, religious gatherings, or others. The five methods were: logistic regression, support vector machines, K nearest

neighbor classification, random forest and gradient boosting. Two oversampling techniques were used, SMOTE and adaptive synthetic (ADASYN) method, which are methods that aim to balance the proportions of deaths/non-deaths. In the end, all five methods were combined with these techniques and for each pair was made an evaluation of its performance. The best AUC was obtained when logistic regression was combined with SMOTE with a value of 0.83.

Goodacre et al. [19] goal was to develop a score that could be used as a triage tool to identify suspected cases of COVID-19. The missing values were solved by using three approaches: the first employed only complete cases, the second approach was based on multiple imputation using chained equations and finally the last used deterministic imputation with missing predictor data assumed to be normal. A multivariate regression with LASSO was performed where the outcome was death or organ support, first with no predictor restrictions and then with the restriction of picking a maximum of ten variables. The selected predictors were the ones incorporated in all three models (age, sex, respiratory rate, systolic BP, oxygen saturation/inspired oxygen ratio, history of renal impairment, performance status, consciousness and respiratory distress). Before developing the score, the research team reviewed these selected variables by considering their clinical practicality and meaning, and decided to remove history of renal impairment and respiratory distress and to add temperature and heart rate because they were recorded as standard procedure. Age categories were created according to their relation with the outcome. Other predictor variables (respiratory rate, heart rate, oxygen saturation, inspired oxygen, systolic BP, consciousness and temperature) were categorised based on an early warning score (NEWS2) as to be easier to assign integer scores. A logistic regression was then performed on the categorised predictors. As the difference between the coefficients and the assigned NEWS2 scores weren't significant, the NEWS2 scores we're used in the respective categories. The remaining predictor scores were based on the model coefficients. The ROC curve was then used to evaluate the score's performance and the AUC was 0.80, compared to the 0.82 in the unrestricted model and 0.81 of the restricted one.

2.5 Imaging techniques

As mentioned earlier, some severe COVID-19 cases suffer from lung difficulties. Therefore, it is fundamental that the techniques used for the analysis of the lungs' condition are useful for predicting the evolution of the disease and from there choose the best ways to treat it.

Computerized Tomography (CT) scans are the best imaging procedure for examination of lung tissues. However, besides being expensive, they require the patient to be moved from nursery or ICU, which increases the risk of transmission and can be very problematic for

patients with severe cases or reduced mobility [20, 21]. Meanwhile ultrasounds, particularly lung ultrasounds, are another imaging technique that has shown some promising results. They are easier to perform as they're faster and don't require patient transport, aside from being less expensive [20].

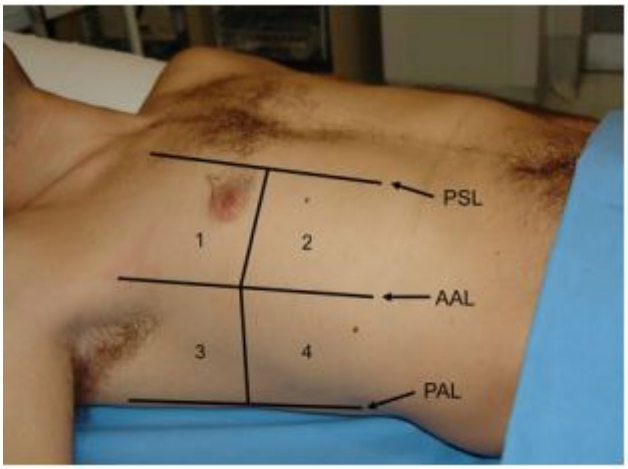

Figure 2.1: Lung division in 4 parts (Zone 1 and 2 are the upper and the lower anterior chest, respectively, and zone 3 and 4 are the upper and the basal lateral chest areas) advised by Volpicelli et al. [22]. PSL parasternal line, AAL anterior axillary line, PAL posterior axillary line. (source: Volpicelli et al. [22])

A common way to analyse lung ultrasounds is by giving scores according to the type of "lines" that appear in different lung zones. The division of the lung is not consensual, Volpicelli et al. [22] recommend dividing each lung in 4 examination zones, the upper and lower anterior chest areas and the upper and basal lateral chest areas, as can be seen in Figure 2.1. De Alencar et al. [23] use 6 zones, considering also the upper and lower posterior zones and Tierney et al. [24] uses 4 zones in the left lung and 5 in the right considering that the right has a greater total lung capacity. Regarding the scores, according to Soldati et al. [25], the classification should be made as follows:

- 0 (normal) - if the ultrasound shows a consistent pleural line and horizontal artifacts (A-lines), Figure 2.2
- 1 - if the pleural line is indented, there are a few vertical lines (B-lines) and small white zones, Figure 2.3
- 2- if the pleural line is broken, the white zones are larger and with more B-lines and some consolidation areas appear, Figure 2.4
- 3 - if the pleural line is broken and there are confluent B-lines with large white areas with or without consolidation areas, Figure 2.5

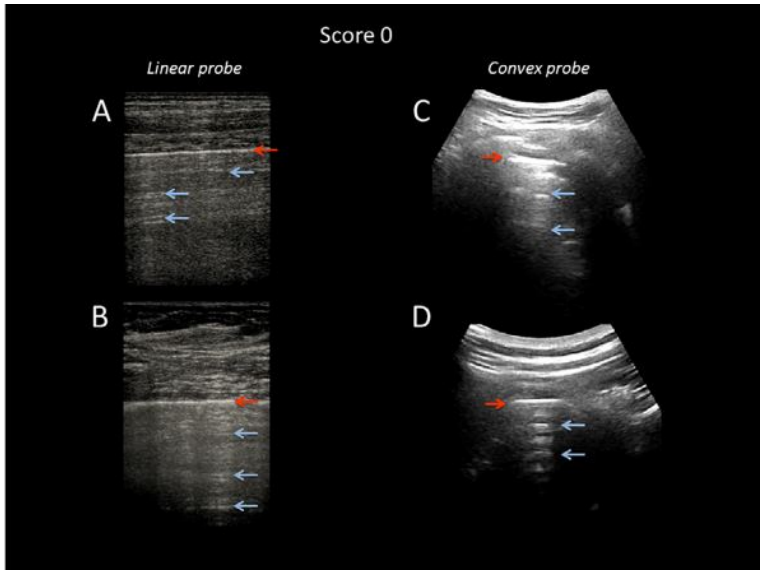

Figure 2.2: Example of Score 0 - the pleural line is continuous (red arrow) and there are horizontal artifacts, the A-lines (blue arrows), (source: Soldati et al. [25]).

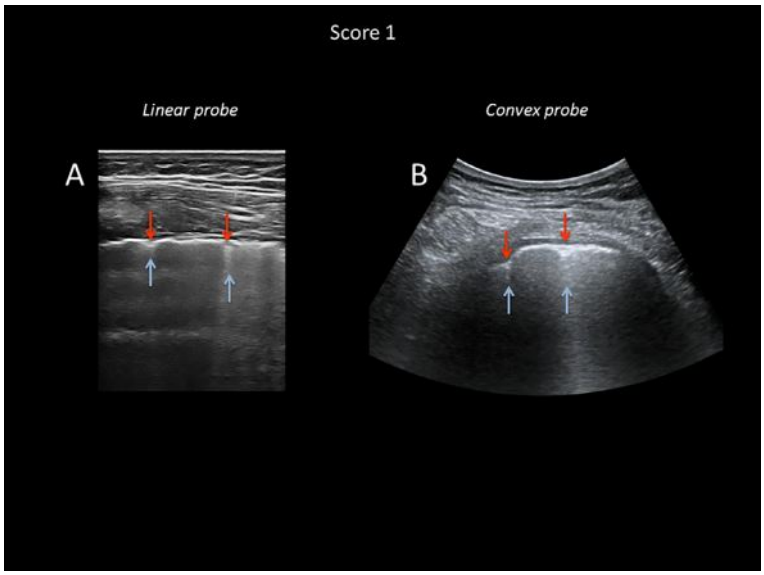

Figure 2.3: Example of Score 1 - the pleural line is not continuous (red arrow) and there are some vertical artifacts, the A-lines (blue arrows), (source: Soldati et al. [25]).

In most cases, after the assessment of these scores, a total score is obtained by making the sum of the individual zones' scores, an approach that has achieved good results. In Deng et al. [21] this type of approach showed a great correlation with CT scores and enabled to distinguish critical-type patients from severe-type patients with very high sensitivity, 97.4%, and specificity, 75.0%, recording an AUC of 0.950. De Alencar et al. [23] also showed promising results by revealing the relation between its version of the Lung Ultrasound scores and all cause mortality (AUC of 0.719). This version was also related with the need for endotracheal intubation (AUC of 0.760) and with ICU admission (AUC of 0.716). Trias-Sabia et al. [26] also found Lung Ultrasound Score (LUS) to be a good

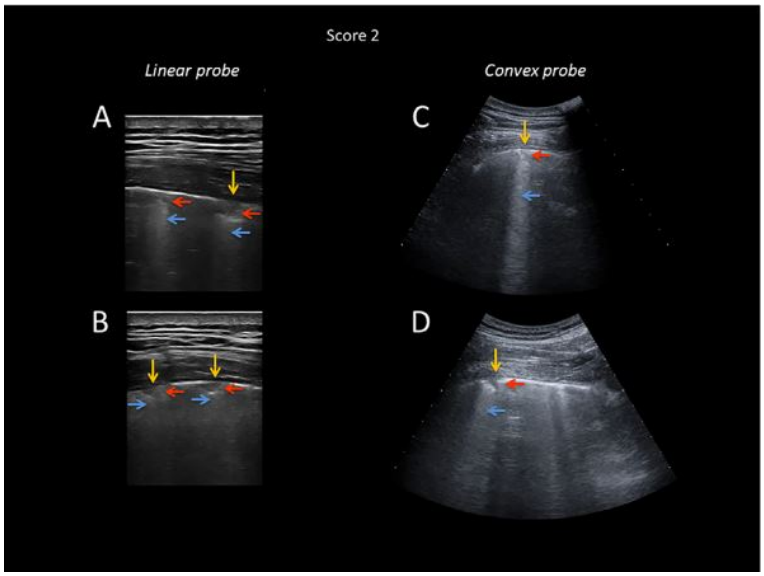

Figure 2.4: Example of Score 2 - the pleural line severely broken, there are white areas (blue arrows) and some small consolidations (dark areas, red arrows), (source: Soldati et al. [25]).

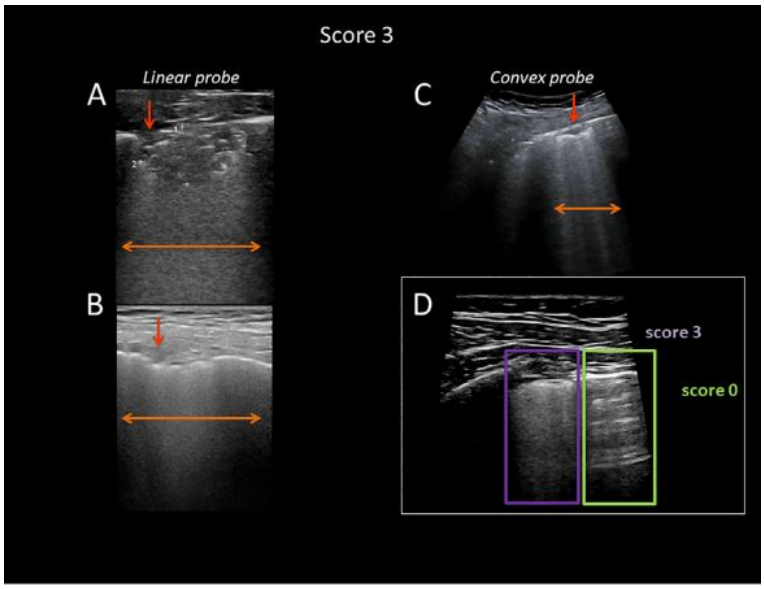

Figure 2.5: Example of Score 3 - the pleural line is severely broken and there are large white areas (orange arrows), (source: Soldati et al. [25]).

predictor for ICU admission or death (AUC of 0.85), besides being correlated with other significant clinical variables. There are also some variation of these scores like Tierney et al. [24] that assigned an extra point for each zone that showed effusion.

In the scope of these encouraging results there have also been developments in the automatic detection of structures in ultrasounds. Baloescu et al. [27] have developed a deep learning algorithm to detect the presence or absence of B-lines and also, in case of presence of these lines, to quantify its severity. Bagon et al. [28] also created a deep

learning algorithm to detect the pleural line and eventual B-lines and then assign a score to the ultrasound to try to predict COVID-19 severity. Chen et al. [29] applied multilayer fully connected neural networks, support vector machine and decision trees to score LUS based on the referred criteria.

2.6 Conclusion

There are several statistical approaches that allow health professionals to make more informed decisions. Composite indicators, GLM's and machine learning models have proved that these techniques can help in accelerating the prediction of outcomes, which can lead to more efficient medical work environments. Lung ultrasounds, besides being a promising tool in evaluating patients with respiratory diseases, like COVID-19, are also already being analysed by these statistical approaches. Therefore, the joint application of GLM's and lung ultrasounds can lead to practical tools for medical use.

3

Data and Methods

3.1 Dataset description

The dataset was made available by Hospital Garcia de Orta, in Almada, under the protocol of the project "Statistical Models for COVID-19 Severity"between Hospital Garcia de Orta and the Center of Mathematics and Applications (CMA) of Mathematics Department of Nova School of Science and Technology. This study work was approved by the Ethics Committee of Hospital Garcia de Orta. It has data about patients with COVID-19 admitted in the hospital from 29/05/2020 to 14/09/2020 in a total of 51. It has sociodemographic information about eventual comorbidities and clinic exams made on the hospital. It has also the scores, for each lung zone, from lung ultrasounds made by these patients, alongside with cardiac and x-Ray complementary data.

3.1.1 Demographics and hospitalisation

The first information about each patient are:

- Age (years)
- Gender - 0 - male; 1 - female
- Date of Admission
- Date of Diagnosis
- Date of ICU hospitalization (if applicable)
- Date of ICU discharge (if applicable)
- Date of hospital discharge
- Death - 0 - no; 1 - yes
- Hospital Length of Stay (days)
- ICU Length of Stay (days)

The average age of the patients was 62.5 years, with 23 males (45%) and 28 females (55%). Of the 51 patients, 31 (61%) were hospitalized in ICU and 8 (16%) eventually died.

3.1.2 Lung ultrasound

The fundamental information of the dataset is in regard to lung ultrasounds. Each patient made more than one lung ultrasound, in different days, which allows to see the patient's health evolution through the analysis of the different ultrasounds . This diagnostic technique measures effusion, with 0 being absence and 1 meaning presence, and then divides the lung in 4 zones, as shown in figure 3.1 , and in each one there are four variables evaluated:

- B-lines - scored from 0 to 3 it is related to vertical lines under the pleural line (visually described in Figures 2.2, 2.3, 2.4 and 2.5):
  - 0 - only A-lines (horizontal) are present and represents a normal lung;
  - 1 - some B-lines (vertical) start to appear with small white zones;
  - 2 - B lines are more common and larger white areas;
  - 3 - Confluent B-lines with large white areas.
- Pleura - 0, when it is homogeneous (normal) and 1, when heterogeneous.
- Subpleural Consolidations (SubP) - 0, when absent and 1, when present.
- Lobar Consolidations (Lob) - 0, when absent and 1, when present.

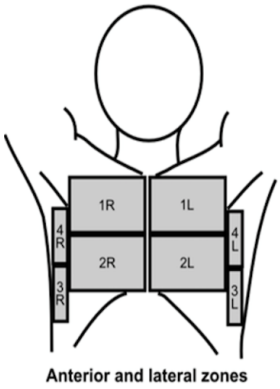

Figure 3.1: Lung division made in the hospital (Zone 1 and 2 are the upper and the lower anterior chest, respectively, and zone 3 and 4 are the basal and the upper lateral chest areas)

3.2 Bootstrap

Bootstrap is a resampling technique used when there is a small original sample and/or when one doesn't know the distribution of a certain statistic $\delta(\mathbf{X}) = \delta(X_1, X_2, ..., X_n)$, where $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ is the original sample of size n.

The basic principle is to create a large number of simulated samples and there are two types of bootstrap, parametric and non-parametric. Parametric bootstrap is based on the assumption that the sample follows a certain distribution, then estimates the parameter and the resampling is made from the distribution with the estimated parameter. The non-parametric bootstrap is built from the original sample and makes resampling with replacement from the original sample.
Let $\boldsymbol{S}_i = (S_{i,1}, S_{i,2}, ..., S_{i,n})$ be the i-th simulated sample obtained from resampling $\boldsymbol{X}$, with $i \in \{1, ..., m\}$ where m is the total number of simulated samples, usually above 1000. For each of these simulated samples it is possible to estimate $\delta_i = \delta(\boldsymbol{S_i})$ and create a bootstrap sample $\boldsymbol{B} = (\delta_1, \delta_2, ..., \delta_m)$ of δ's distribution. it is now possible to estimate the mean (μ_B), the standard deviation (σ_B) and the quantile of probability α ($q_{\boldsymbol{B},\alpha}$) of $\boldsymbol{B}$.
Usually, the point estimate for δ is $\mu_{\boldsymbol{B}}$ and there are two different confidence intervals that can be constructed:

- Normal Confidence interval

$$[\mu_{\boldsymbol{B}} - \sigma_{\boldsymbol{B}} \times z_{1-\alpha/2}, \mu_{\boldsymbol{B}} + \sigma_{\boldsymbol{B}} \times z_{1-\alpha/2}]$$

  where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

- Percentile Confidence Interval

$$[q_{\boldsymbol{B},\alpha/2}, q_{\boldsymbol{B},1-\alpha/2}].$$

3.3 Goodness of fit

To assess the goodness of fit of any distribution to certain data goodness of fit tests are a valuable asset to compare the data's distribution to the known theoretical distributions. The theory of this subsection is based on Conover [30].

3.3.1 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (KS) test, as is more commonly known, is a goodness of fit test that can perform two types of comparisons. The first is to compare the sample to a theoretical distribution, the one explored in this section, and the second is to compare two samples between them to check if they may have the same distribution and is based on the vertical distance between the empirical distribution function $S(x)$ and some distribution function, $F^*(x)$.
Let $(X_1, X_2, ..., X_n)$ be a random sample, with some unknown distribution function, denoted by $F(x)$. The empirical distribution function $S(x)$ is a function of x, which equals the fraction of X_i's that are less or equal to x for each $x \in (-\infty < x < \infty)$.

The test statistic, for the two side test, is the supremum of the vertical distance between $S(x)$ and $F(x)$:

$$T = sup_x|F^*(x) - S(x)|. \tag{3.1}$$

This vertical distance T can be divided in two, T^+ and T^-, used for one sided tests that can be defined as:

$$\begin{aligned} T^+ &= sup_x[F^*(x) - S(x)] \\ T^- &= sup_x[S(x) - F^*(x)]. \end{aligned} \tag{3.2}$$

The null distribution, i.e. the distribution when the null hypothesis is true, of T^+ and T^- is given by:

$$G(x) = 1 - x \sum_{i=0}^{[n(1-x)]} \left(\frac{n}{i}\right)\left(1 - x - \frac{i}{n}\right)^{n-i}\left(x + \frac{i}{n}\right)^{i-1}. \tag{3.3}$$

Where, $[n(1-x)]$ is the greatest integer less than or equal to $n(1-x)$. As T is less then x only when both T^+ and T^- are less than x, the approximate distribution function of T is:

$$P(T \leq x) = P(T^+ \leq x \wedge T^+ \leq x) = [G(x)]^2. \tag{3.4}$$

The hypothesis can than be written as:

$$\begin{aligned} H_0 &: F(x) = F^*(x) \\ H_1 &: F(x) \neq F^*(x). \end{aligned} \tag{3.5}$$

The rejection of the null hypothesis, at the level of significance α, is made if T exceeds the $1-\alpha$ quantile tabled (See table A13 of [30]). Some limitations of KS test are the fact that it can only be applied to continuous distributions and it is far more sensitive in the central values than in the tails.

3.3.2 Shapiro-Wilks Normality test

The Shapiro-Wilks Normality test is also a non-parametric goodness of fit test used to compare the sample's distribution with the normal distribution.
Let$(X_1, X_2, ..., X_n)$ be a random sample associated with an unknown distribution function $F(x)$. To calculate the test statistic some intermediate variables must be computed. Let D be:

$$D = \sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{3.6}$$

where $\bar{X}$ is the sample mean. Let $X^{(1)} \leq X^{(2)} \leq ... \leq X^{(n)}$ denote the order statistics. From Table A16 of [30], for the observed sample size n, obtain the coefficients $b_1, ..., b_k$, where k

is approximately $n/2$. The test statistic T_3 is given by:

$$T_3 = \frac{1}{D}\Big[\sum_{i=1}^{k} b_i(X^{(n-i+1} - X^{(i)})\Big]^2. \tag{3.7}$$

The test statistic T_3 is basically the square of a correlation coefficient, where the Pearson correlation coefficient is computed between the order statistics $X^{(i)}$ in the sample and the scores b_i, which represent what the order statistics should look like if the population is normal. Thus if T_3 is close to 1.0 the sample behaves like a normal sample. If T_3 is too small, that is, too far below 1.0, the sample looks non-normal. The hypotheses can be written as:

$$\begin{aligned} H_0 &: F(x) \text{ is a normal distribution function} \\ H_1 &: F(x) \text{ is non-normal} \end{aligned} \tag{3.8}$$

H_0 will be rejected at the level of significance α if T_3 is less than the α quantile tabled.

3.4 Generalized Linear Models

Generalized Linear Models (GLM) were introduced by Nelder and Wedderburn in 1972 [31] and are a generalization of the classic linear regression model. While linear regression is a great tool to model some situations, there are a lot of cases where it is not the adequate solution. For example, if one wants to model a certain probability then linear regression may lead to values lower then 0 and higher than 1. In these cases, other GLM's are more appropriate, as certain transformations may lead to the desired limits. The following sections are based on Turkman, et al. [32].

3.4.1 Linear Regression

Regression models are based on the idea that the value of a certain variable, called independent or response variable, can be explained using the values of other observed variables, called independent or explanatory. The simplest regression is the linear regression which states that each explanatory variable contributes in a linear way to the response, assuming normal random residuals.
Let $(X_1, X_2, ..., X_m)$ be the m explanatory variables of the model and let n be the number of observations of the sample. Let $x_{i,j}$ represent the i-th observation of the j-th explanatory variable. The response variable, as a random variable, is denoted by Y_i while the realization of that variable is expressed by y_i. This model has some important assumptions about the residuals: independent and normally distributed residuals with constant variance. That being said the structure of the model is:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_m x_{i,m} + \epsilon_i \tag{3.9}$$

where $\beta_0,...,\beta_m$ are the parameters of the model and ϵ_i's are the residuals, that must follow the assumptions, i.e,

$$\epsilon_i \sim N(0,\sigma^2), \text{ independent.} \tag{3.10}$$

Let $\boldsymbol{\mu} = (\mu_1,...,\mu_n)$ be defined as:

$$\mu_i = E[Y_i|\boldsymbol{x_i} = (x_{i,1}, x_{i,2},..., x_{i,m})] = \beta_0 + \beta_1 x_{i,1} + ... + \beta_m x_{i,m} \tag{3.11}$$

an approach that will be useful to better understand generalized linear models.
The estimation of $\boldsymbol{\beta} = (\beta_0, \beta_1,..., \beta_m)$, whose estimate will be represented by $\hat{\boldsymbol{\beta}}$, is made with the goal of having the model's estimated fitted value adjusted $\hat{\mu_i} = \hat{\beta_0} + \hat{\beta_1} x_{i,1}... + \hat{\beta_m} x_{i,m}$ as close as possible of the real value y_i. To solve this problem the least squares method is used where the aim is the minimization of the sum of the squares of the errors:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \mu_i)^2 = \min_{\beta} \sum_{i=1}^{n} r_i^2. \tag{3.12}$$

3.4.1.1 Estimation

The least squares aim can be rewritten in matricial form in order to be easier to see how one can estimate $\boldsymbol{\beta}$.

$$\min_{\beta} \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \mu_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1} + ... + \beta_m x_{i,m}))^2. \tag{3.13}$$

Defining X, $\boldsymbol{\beta}$ and $\boldsymbol{y}$ as:

$$X = \begin{bmatrix} 1 & x_{1,1} & ... & x_{1,m} \\ 1 & x_{2,1} & ... & x_{2,m} \\ ... & ... & ... & ... \\ 1 & x_{n,1} & ... & x_{n,m} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_m \end{bmatrix} \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix}.$$

Thus:

$$\boldsymbol{\mu} = E[Y|X] = X\boldsymbol{\beta} = \begin{bmatrix} 1 & x_{1,1} & ... & x_{1,m} \\ 1 & x_{2,1} & ... & x_{2,m} \\ ... & ... & ... & ... \\ 1 & x_{n,1} & ... & x_{n,m} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_m \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{1,1} + ... + \beta_M x_{1,m} \\ \beta_0 + \beta_1 x_{2,1} + ... + \beta_M x_{2,m} \\ ... \\ \beta_0 + \beta_1 x_{n,1} + ... + \beta_M x_{n,m} \end{bmatrix}.$$

And finally 3.13 can be written in matricial form as:

$$\begin{aligned} \min_{\beta} \sum_{i=1}^{n} (y_i - \mu_i)^2 &= \min_{\beta} (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta}) \\ &= \min_{\beta} (\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T X\boldsymbol{\beta} - \boldsymbol{\beta}^T X^T \boldsymbol{y} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}) \\ &= \min_{\beta} SSR. \end{aligned} \tag{3.14}$$

As the minimization is in regard to $\boldsymbol{\beta}$ then equalling the derivative, with respect to $\boldsymbol{\beta}$, to 0 will lead to the desired values.

$$\frac{\partial_{SSR}}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{y}^T X\boldsymbol{\beta} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}) = -2X^T y + 2X^T X\boldsymbol{\beta} = 0. \tag{3.15}$$

Which ultimately leads to:

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}. \tag{3.16}$$

Although this is only a necessary condition, for the sufficient condition will need for the second derivative, Hessian matrix, to be definite positive.

3.4.2 Exponential family

The exponential family is a group of distributions whose probability density function can be written in the form:

$$f(x) = \exp\left(\frac{x\theta - b(\theta)}{a(\phi)} + c(x,\phi)\right). \tag{3.17}$$

Where θ and ϕ are scalar parameters, also known as the location and dispersion parameter, respectively, and a(),b() and c() are known real functions.
Some examples of distributions that belong to the exponential family are:

- Normal Distribution - $N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \exp\left(\frac{x\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\left(\frac{x^2}{\sigma^2} + \ln\left(2\pi\sigma^2\right)\right)\right) \tag{3.18}$$

  where:

  - $\theta = \mu$
  - $b(\theta) = \frac{\mu^2}{2}$
  - $a(\phi) = \phi = \sigma^2$
  - $c(x,\phi) = -\frac{1}{2}\left(\frac{x^2}{\sigma^2} + \ln\left(2\pi\sigma^2\right)\right)$.

- Gamma Distribution - $G(\alpha, \beta)$

$$\begin{aligned} f(x) &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \\ &= \exp\left(\alpha \ln(\beta) - \ln(\Gamma(\alpha)) + (\alpha-1)\ln(x) - \beta x\right) \\ &= \exp\left(\frac{-\frac{\beta}{\alpha}x + \ln(\beta)}{\frac{1}{\alpha}} + (\alpha-1)\ln(x) - \ln(\Gamma(\alpha))\right). \end{aligned} \tag{3.19}$$

By making the change of variables $\alpha = \frac{1}{\phi}$ and $\beta = \frac{\theta}{\phi}$ the expression becomes:

$$f(x) = \exp\left(\frac{\theta x - \ln(\theta)}{-\phi} + \frac{\ln(\theta)}{\theta} + \left(\frac{1}{\phi} - 1\right)\ln(x) - \ln\left(\Gamma\left(\frac{1}{\phi}\right)\right)\right). \tag{3.20}$$

Where:

- $\theta = \frac{\beta}{\alpha}$
- $b(\theta) = \ln\left(\frac{\beta}{\alpha}\right)$
- $a(\phi) = -\frac{1}{\alpha}$
- $c(x,\phi) = \frac{\ln(\theta)}{\theta} + \left(\frac{1}{\phi} - 1\right)\ln(x) - \ln\left(\Gamma\left(\frac{1}{\phi}\right)\right)$.

3.4.3 Maximum likelihood

Maximum likelihood is the most popular technique for parameter estimation. The main idea is that the Maximum Likelihood Estimator (MLE) is the parameter point for which the observed sample is the most likely and the essential tool is the likelihood function. This subsection is based on Casella, et al. [33].
Let $\boldsymbol{X} = (X_1,...,X_n)$ be an independent and identically distributed sample from a population with probability density function given by $f(x|\boldsymbol{\theta} = (\theta_1,...,\theta_k))$, for some $\boldsymbol{\theta}$ in the parameter space Θ. The likelihood function is defined by:

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}). \tag{3.21}$$

For each sample point $\boldsymbol{x}$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ be a parameter value at which $L(\boldsymbol{\theta}|\boldsymbol{x})$ attains its maximum as a function of $\boldsymbol{\theta}$, with $\boldsymbol{x}$ held fix. A MLE of the parameter $\boldsymbol{\theta}$ based on sample $\boldsymbol{X}$ is $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$.
If the Likelihood function is differentiable (in θ_i), then possible candidates for MLE are the values of $(\theta_1,...,\theta_k)$ that solve:

$$\frac{\partial}{\partial\theta_i} L(\boldsymbol{\theta}|\boldsymbol{x}) = 0; i = 1,...,k. \tag{3.22}$$

It is important to note that this is only a necessary condition, not a sufficient one. In practice, instead of using the likelihood function, the logarithm of the likelihood function, $l(\boldsymbol{\theta}|\boldsymbol{x})$ is used as the logarithm is an increasing function and therefore the maximum of the likelihood function is obtained at the same $\boldsymbol{\theta}$ of the maximum of the log-likelihood and in many cases it is easier to differentiate.

3.4.3.1 Likelihood Ratio Test

The likelihood ratio test statistic for testing:

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \text{ versus } \boldsymbol{\theta} \in \Theta_0^C \tag{3.23}$$

where, $\Theta_0 \in \Theta$ and $\Theta_0 \cup \Theta_0^C = \Theta$, is given by:

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\Theta_0} L(\boldsymbol{\theta}|\boldsymbol{x})}{\sup_{\Theta} L(\boldsymbol{\theta}|\boldsymbol{x})} \tag{3.24}$$

where $\lambda(\boldsymbol{x})$ is between 0 and 1. The Wilks' statistic is defined by:

$$\begin{aligned} \Lambda &= -2ln(\lambda(\boldsymbol{x})) \\ &= -2(\sup_{\Theta_0} l(\boldsymbol{\theta}|\boldsymbol{x}) - \sup_{\Theta} l(\boldsymbol{\theta}|\boldsymbol{x})) \\ \Lambda &\overset{a}{\sim} \chi_q^2 \end{aligned} \tag{3.25}$$

where q is the difference in dimensionality between Θ and Θ_0.

3.4.4 Modeling

The GLM's are a generalization of the linear regression where the distribution of the response variable belongs to the exponential family (not only the Normal distribution). The idea is that instead of using the linear predictor, $\eta_i = X_i\boldsymbol{\beta}$, $X_i = (x_{1i},...,x_{ni})$ to estimate the response, $\mu_i = \eta_i$, the value of μ_i is a function of the linear predictor, $\mu_i = h(\eta_i)$. This function, $h()$, must be monotone and differentiable and its inverse, $h^{-1}() = g()$, is called the link function as $g\left(\mu_i\right) = \eta_i$. This link functions are chosen according to the response variable's distribution.
Some examples of link functions, with an example of the distributions where they may applied, are:

- identity - $g(\mu) = \mu$ - Gaussian
- reciprocate $g(\mu) = \frac{1}{\mu}$ - Gamma
- inverse quadratic $g(\mu) = \frac{1}{\mu^2}$ - Wald
- square root $g(\mu) = \sqrt{\mu}$ - Poisson
- logarithmic $g(\mu) = ln(\mu)$ - Gamma
- logit $g(\mu) = ln\left(\frac{\mu}{1-\mu}\right)$ - Binomial

3.4.5 Parameter estimation

In generalized linear models the key tool to determine the parameters is maximum likelihood estimation. Defining X, $\boldsymbol{\beta}$ and $\boldsymbol{y}$, again, as:

$$X = \begin{bmatrix} 1 & x_{1,1} & ... & x_{1,m} \\ 1 & x_{2,1} & ... & x_{2,m} \\ ... & ... & ... & ... \\ 1 & x_{n,1} & ... & x_{n,m} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_m \end{bmatrix} \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix}.$$

The linear predictor is transformed through $h()$:

$$\eta = X\beta = \begin{bmatrix} 1 & x_{1,1} & ... & x_{1,m} \\ 1 & x_{2,1} & ... & x_{2,m} \\ ... & ... & ... & ... \\ 1 & x_{n,1} & ... & x_{n,m} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_m \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{1,1} + ... + \beta_M x_{1,m} \\ \beta_0 + \beta_1 x_{2,1} + ... + \beta_M x_{2,m} \\ ... \\ \beta_0 + \beta_1 x_{n,1} + ... + \beta_M x_{n,m} \end{bmatrix}.$$

And $\mu = h(\eta)$.
As was referred in 3.17, the probability density function of a distribution belonging to the exponential family can be written as:

$$f(y|\theta,\phi) = \exp\left(\frac{\theta y - b(\theta)}{a(\phi)} + c(y,\phi)\right) \tag{3.26}$$

where θ is the location parameter and ϕ is the dispersion parameter.
However in some cases, like the normal and gamma distributions shown earlier, $a(\phi) = \frac{\phi}{\omega}$ where ω is a constant. Therefore one can also write:

$$f(y|\theta,\phi,\omega) = \exp\left(\frac{\omega}{\phi}(\theta y - b(\theta)) + c(y,\phi)\right). \tag{3.27}$$

It is important to note that θ_i is a function of β, as $b'(\theta_i) = \mu_i = h(\eta_i)$, and $\eta_i = X\beta$.
The likelihood function can, therefore, be defined as a function of β, which for simplification will be written as $L(\beta)$:

$$\begin{aligned} L(\beta,\phi,\omega_i|\mathbf{y}) &= \prod_{i=1}^{n} f(y_i|\theta_i,\phi,\omega_i) = \\ &= \prod_{i=1}^{n} exp\left(\frac{\omega_i}{\phi}(\theta_i y_i - b(\theta_i)) + c(y_i,\phi)\right) \\ &= \exp\left(\frac{1}{\phi}\sum_{i=1}^{n}\omega_i(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(y_i,\phi,\omega_i)\right) \\ &= L(\beta) \end{aligned} \tag{3.28}$$

and therefore log-likelihood is:

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \frac{\omega_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y_i,\phi,\omega_i) \\ &= \sum_{i=1}^{n} l_i(\beta). \end{aligned} \tag{3.29}$$

To find the maximum likelihood estimators the following system must be solved:

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_j} = 0, j = 0, ..., m \\ \frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_j} &= \frac{\partial l_i(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i(\mu_i)}{\partial \mu_i} \frac{\partial \mu_i(\eta_i)}{\partial \eta_i} \frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \beta_j}.\end{aligned} \tag{3.30}$$

Where:

- $\frac{\partial l_i(\theta_i)}{\partial \theta_i} = \frac{\omega_i(y_i - \mu_i)}{\phi}$
- $\frac{\partial \theta_i(\mu_i)}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \frac{\phi}{\omega_i Var(Y_i)}$ (a result from the exponential family)
- $\frac{\partial \eta_i(\beta)}{\partial \beta_j} = x_{i,j}$

and consequently:

$$\frac{\partial l_i(\beta)}{\partial \beta_j} = \frac{\omega_i(y_i - \mu_i)}{\phi} \frac{\phi}{\omega_i Var(Y_i)} \frac{\partial \mu_i(\eta_i)}{\partial \eta_i} x_{i,j} = \frac{(y_i - \mu_i) x_{i,j}}{Var(Y_i)} \frac{\partial \mu_i(\eta_i)}{\partial \eta_i} \tag{3.31}$$

$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ is also called the score function of $\boldsymbol{\beta}$, $s(\boldsymbol{\beta})$, and similarly $\frac{\partial l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}) = s_i(\boldsymbol{\beta})$.
The covariance matrix of the score function is called Fisher information matrix, which will be central to the estimation of $\boldsymbol{\beta}$, it is defined as $I(\boldsymbol{\beta}) = E\Big[- \frac{\partial s(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Big]$, which require the calculation of the second derivatives of the log-likelihood:

$$I(\boldsymbol{\beta}) = E\left[- \frac{\partial s(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right] = E\left[- \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \right]. \tag{3.32}$$

Therefore:

$$I(\boldsymbol{\beta})_{j,k} = \sum_{i=1}^{n} E\left[- \frac{\partial^2 l_i(\boldsymbol{\beta})}{\partial \beta_j \beta_k} \right] = \sum_{i=1}^{n} \frac{x_{i,j} x_{i,k}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \tag{3.33}$$

and if one wants the matricial form:

$$I(\boldsymbol{\beta}) = X^T W X \tag{3.34}$$

where W is a diagonal matrix where the i-th diagonal value is:

$$\frac{\left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2}{Var(Y_i)}. \tag{3.35}$$

3.4.6 Iterative pondered least squares method

This method uses the score function, the Fisher Information matrix and a method to solve non-linear equations to estimate β. The starting point is the expression:

$$\hat{\beta}^{k+1} = \hat{\beta}^k + \left[I(\hat{\beta}^k) \right]^{-1} s(\hat{\beta}^k). \tag{3.36}$$

where $\hat{\beta}^k$ is the k-th iteration of the method, $I()$ is the Fisher Infomation matrix and $s()$ is the score function. This method is very similar to Newton-Rapson, the referred method to solve non-linear equations, but instead of using the Hessian matrix it uses the Fisher Information. This allows a smaller computation effort as Fisher Information is easier to calculate and Fisher Information is always definite positive, which is an advantage for this type of iterative methods.
The above expression can be rewritten as:

$$\left[I(\hat{\beta}^k) \right] \hat{\beta}^{k+1} = \left[I(\hat{\beta}^k) \right] \hat{\beta}^k + s(\hat{\beta}^k). \tag{3.37}$$

From equations (3.31) and (3.33) the l-th element of the right side of (3.37) is:

$$\sum_{j=1}^{p} \left[\sum_{i=1}^{n} \frac{x_{i,j} x_{i,l}}{Var(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \right] \beta_j^k + \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{i,l}}{Var(y_i)} \frac{\partial \mu_i(\eta_i)}{\partial \eta_i}. \tag{3.38}$$

And therefore in the matricial form:

$$\left[I(\hat{\beta}^k) \right] \hat{\beta}^{k+1} = X^T W^k \boldsymbol{u}^k. \tag{3.39}$$

Where $u_i^k = \sum_{j=1}^{p} x_{i,j} \beta_j^k + (y_i - \mu_i^k) \frac{\partial \eta_i^k}{\partial \mu_i^k}$ and W^k is the matrix defined in (3.35). And finally the algorithm can be expressed in the matricial form as:

$$\hat{\beta}^{k+1} = (X^T W^K X)^{-1} X^T W^k \boldsymbol{u}^k. \tag{3.40}$$

3.4.7 Model validation and selection

So far it has been explained how GLM can be developed. However, it was assumed that all explanatory variables were meaningful to the value of the response. This is not always the case as some variables may not have a significant influence and the goal is not only for the model to best fit the data but also to be only as complex as it needs to be, i.e. for the model to be parsimonious. For this purposes a number of statistical tools were developed.

Before entering in detail in each method it is important to define models:

- Saturated model

The saturated model is defined as one that has as many parameters as the number of points it fits and therefore every predicted value is equal to the observed one. This models ignores the linear predictor, η, and assigns linear independent vectors as the explanatory variables for each sample entrance and therefore the prediction will be perfect. It is mainly used as the ideal model and by comparison if a model has similar test results it is a good indicator.

- Null model

  The null model is the one that uses only the intercept, a constant, and therefore the prediction for each data point is the mean of the response variable. It is the simplest model possible and it is used as a starting point for some algorithms and for comparison with other models.

- Maximal model

  The maximal model is the one that uses all the explanatory variables. It is the most complex model and it is also used as a starting point for some algorithms and for comparison with other models.

3.4.7.1 Goodness of Fit statistics

Deviance The deviance (D) is a goodness of fit statistic that compares the saturated model with the ones constructed using the actual explanatory variables. It uses maximum likelihood functions to make comparisons, more precisely using the Wilks' statistic. As mentioned earlier (3.25) the Wilks' statistic is given by the expression:

$$\Lambda = -2(\sup_{\Theta_0} l(\boldsymbol{\theta}|\boldsymbol{x}) - \sup_{\Theta} l(\boldsymbol{\theta}|\boldsymbol{x})). \tag{3.41}$$

When applied to model validation the comparison is made between the log-likelihood of the constructed model, M, and the one from the saturated model, S.

$$D = -2\Big[l(\hat{\beta_M}) - l(\hat{\beta_S})\Big]. \tag{3.42}$$

where $\hat{\beta_M}$ and $\hat{\beta_S}$ are the estimates of the parameters in the constructed and saturated models, respectively.
As shown in (3.29) the log-likelihood of any distribution that belongs to the exponential family can be written as:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\omega_i(y_i q(\mu_i) - b(q(\mu_i)))}{\phi} + c(y_i, \phi, \omega_i) \tag{3.43}$$

where $q(\mu_i) = \theta_i$ is used to highlight the relation between θ_i and μ_i.
Therefore, for the saturated model, as $\hat{\mu_i} = y_i$, the log-likelihood becomes

$$l(\beta) = ln(L(\beta)) = \sum_{i=1}^{n} \frac{\omega_i(y_i q(y_i) - b(q(y_i)))}{\phi} + c(y_i, \phi, \omega_i) \tag{3.44}$$

and the final expression for the deviance is:

$$D = -2\Big[l(\hat{\beta_M}) - l(\hat{\beta_S})\Big] = -2\sum_{i=1}^{n} \frac{\omega_i}{\phi}\Big(\Big[y_i q(\hat{\mu_i}) - b(q(\hat{\mu_i}))\Big] - \Big[y_i q(y_i) - b(q(y_i))\Big]\Big). \tag{3.45}$$

The deviance function can also be used to compare nested models. A model M_1 is said to be nested in M_2 if the set of explanatory variables of M_1 is a subset of the explanatory variables of M_2. By comparing the deviance of the two models it is possible to make an assessment on the importance of the explanatory variables used in M_2 and not in M_1. If the difference is small, then M_1 is a better model as is more parsimonious, otherwise M_2 is preferred. The deviance expression is:

$$\begin{aligned}
\Delta D &= -2\Big[l(\hat{\beta_{M_1}}) - l(\hat{\beta_{M_2}})\Big] \\
&= -2\Big[\Big(l(\hat{\beta_{M_1}}) - l(\hat{\beta_S})\Big) - \Big(l(\hat{\beta_{M_2}}) - l(\hat{\beta_S})\Big)\Big] \\
&= \Big[-2\Big(l(\hat{\beta_{M_1}}) - l(\hat{\beta_S})\Big)\Big] - \Big[-2\Big(l(\hat{\beta_{M_2}}) - l(\hat{\beta_S})\Big)\Big] \\
&= D_{M_1} - D_{M_2}.
\end{aligned} \tag{3.46}$$

Assuming that M_1 has m_1 variables and M_2 has $m_1 + m_2$ variables, i.e. m_1 is the number of common variables and m_2 the number of variables exclusive to M_2, then $D_{M_1} \sim \chi^2_{n-m_1}$ and $D_{M_2} \sim \chi^2_{n-m_1-m_2}$. From there, using the additive property of the χ^2 distribution, $D_{M_1} - D_{M_2} \sim \chi^2_{m_2}$ and it is then possible to obtain the critic values for the test associated with the desired significance.

Pearson's statistic Another Goodness of Fit (GOF) statistic is the Pearson's statistic:

$$X^2 = \sum_{i=1}^{n} \frac{\omega_i(y_i - \hat{\mu_i})^2}{Var(\hat{\mu_i})}. \tag{3.47}$$

It has a more intuitive interpretation, as it is the sum of the squares of the residuals, pondered by ω_i, divided by the variance. A χ^2 approximation can also be established with $n - m$ degrees of freedom, where m is the number of covariates.

Both these methods have their own advantages, the additive property in the deviance and the interpretation of Pearson's statistic, and can be complementary. In the Normal distribution both methods are equal as they represent the sum of the square of the residuals.

Akaike Information Criterion The Akaike Information Criterion (AIC) is another important GOF statistic to compare models because it accounts the number of variables used to fit the data. Assuming a model with coefficients $\hat{\beta}$ and $\hat{\phi}$ as the estimator of ϕ (that can be estimated by the expression $\hat{\phi} = \frac{1}{n-m}\sum_{i=1}^{n}\frac{\omega_i(y_i-\hat{\mu_i})^2}{Var(\hat{\mu_i})}$) , then:

$$AIC = -2l(\hat{\beta}) + 2m \tag{3.48}$$

where m is the size of $(\hat{\beta})$. As can be seen, this method balances the log-likelihood obtained with the estimated parameters with the number of parameters, and will be very useful in the selection algorithms.

Coefficient of determination - R^2 The Coefficient of determination is also a GOF statistic that indicates the relation between the Residual sum of squares (SSR) and the total sum of squares (SST), which is defined as:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y}) \tag{3.49}$$

where $\bar{y}$ is the mean of $\boldsymbol{y}$. The value of R^2 is given by

$$R^2 = 1 - \frac{SSR}{SST}. \tag{3.50}$$

It varies between 0 and 1, with 1 being the perfect fit, as it would mean that $y_i = \hat{\mu i}_i$. It is used to assess the quality of linear regression models.

3.4.7.2 Selection

Model selection consists in choosing the models that best balance goodness of fit and parsimony.

Hypothesis testing It is also possible to compare two nested models M_1 and M_2, where M_1 is the one with less variables. Let $\boldsymbol{\beta} = (\tilde{\boldsymbol{\beta_1}}, \tilde{\boldsymbol{\beta_2}})$ be the variables' coefficients, with $\tilde{\boldsymbol{\beta_1}}$ the coefficients of the m_1 common variables and $\tilde{\boldsymbol{\beta_2}}$ the coefficients of the m_2 variables used in M_2 and not in M_1. Comparing M_1 with M_2 can be made by testing if the coefficients $\tilde{\boldsymbol{\beta_2}}$ are equal to zero. If that is not rejected it leads to the conclusion that the respective variables may not be useful and therefore M_1 is better than M_2 as is more parsimonious. The referred test can be written as:

$$H_0 : \tilde{\boldsymbol{\beta_2}} = 0 \text{ versus } H_1 : \tilde{\boldsymbol{\beta_2}} \neq 0. \tag{3.51}$$

A possible method to materialize this test is the F test.

$$F = \frac{(D_{M_1} - D_{M_2})/(m_2)}{D_{M_2}/(n - m_1 - m_2)} \sim F_{m_2, n-m_1-m_2}. \tag{3.52}$$

If the increase of deviance of M_1 is small in comparison to M_2 then F will be small and therefore the p-value of the test will be high, leading to not rejecting H_0. This would mean that both models had similar performances and therefore M_1 should be chose as has less variables; on the other hand if the increase in the deviance is high, then the p-value is small which means that H_0 is rejected and that the variables subtracted from M_2 to create M_1 are important in describing the response variable.

3.4.8 Algorithms

It has now been seen how models can be compared and the way to quantify this comparisons. However, in data with a considerable number of variables, the number of possible models becomes a combinatory problem and for that reason some selection algorithms were designed to simplify the task of model selection.
Let $\boldsymbol{X} = (X_1, ..., X_m)$ be the explanatory variables and $M(\boldsymbol{V}), \boldsymbol{V} \subseteq \boldsymbol{X}$, represents the model constructed using the explanatory variables in $\boldsymbol{V}$. For example, $M(\boldsymbol{X})$ represents the maximal model and $M(\emptyset)$ is the null model. Let $AIC(M)$ be the function that returns the AIC of the model M, and, for simplification AIC_n represents the AIC of the new model, while AIC_o represents the equivalent but to the old model. Finally $pv(x)$ is the value of the p-value of the significance test of the x variable in a given model.

Forward stepwise The initial model in this selection method is the null model and the idea is to start adding the more significant variables until reaching the stop condition. In this case the stop condition will be when the AIC is not improved or when there are no more variables to include. AIC_o is initialized as a big value just to make sure the stop condition isn't verified before starting the actual selection.

Backward stepwise Similarly it is possible to make selection thinking in the inverse order. Instead of starting with the null model, the first model is the maximal, and instead of adding the most significant variable the least significant is withdrawn, as long as doesn't deteriorate the goodness of fit of the model (AIC).

Bidirectional stepwise This final method is a combination of the previous two. It starts with the null model and starts adding variables, just like in forward selection, however at each iteration, after adding a variable, it tests the effect of withdrawing one of the variable from the model, like in backward selection.
Starting from these methods a lot of subtle differences can be applied, whether it is on the stop criteria where instead of AIC one can choose a threshold for the p-value test and variables are only added while its significance are under the threshold; or the choosing of the variable to enter the model, instead of the best p-value one can choose the variable that most improves AIC.

Algorithm 1 Forward Stepwise

$V_c \leftarrow \emptyset$ ▷ Variables in the current model
$V_m \leftarrow \emptyset$ ▷ Variables in the best model
$X \leftarrow \{X_1, ..., X_p\}$ ▷ Variables
$AIC_n \leftarrow AIC(M(V_c))$ ▷ AIC of the new model
$AIC_o \leftarrow 100000$ ▷ AIC of the old model
while $AIC_n < AIC_o$ **do**
  $V_m \leftarrow V_c$ ▷ The best model is the current model
  $V_r \leftarrow X \backslash V_c$ ▷ The variables not in the current model
  $best \leftarrow 1, add \leftarrow \emptyset$ ▷ Initialize best and add
  for x in V_r **do**
    $p \leftarrow pv(x)$ in $M(x \cup V_c)$
    if $p < best$ **then** ▷ Check the best p-value
      $best \leftarrow p$
      $add \leftarrow x$ ▷ Save the variable with best p-value
    end if
  end for
  $V_c \leftarrow V_c \cup add$ ▷ Add the best variable to V_c
  $AIC_o \leftarrow AIC_n$ ▷ Update AIC's
  $AIC_n \leftarrow AIC(M(V_c))$
end while

Algorithm 2 Backward Stepwise

$X \leftarrow \{X_1, ..., X_p\}$ ▷ Variables
$V_c \leftarrow X$ ▷ Variables in the current model
$V_m \leftarrow X$ ▷ Variables in the best model
$AIC_n \leftarrow AIC(M(V_c))$ ▷ AIC of the new model
$AIC_o \leftarrow 100000$ ▷ AIC of the old model
while $AIC_n < AIC_o$ **do**
  $V_m \leftarrow V_c$ ▷ The best model is the current model
  $worst \leftarrow 0, sub \leftarrow \emptyset$ ▷ Initialize worst and sub
  for x in V_c **do**
    $p \leftarrow pv(x)$ in $M(V_c)$
    if $p > worst$ **then** ▷ Check the worst p-value
      $worst \leftarrow p$
      $sub \leftarrow x$ ▷ Save the variable with worst p-value
    end if
  end for
  $V_c \leftarrow V_c \backslash add$ ▷ Subtract the worst variable to V_c
  $AIC_o \leftarrow AIC_n$ ▷ Update AIC's
  $AIC_n \leftarrow AIC(M(V_c))$
end while

3.4.9 Analysis of residuals

The residuals are an important part of every model as they give information about the effect of previous choices (distribution of the response variable, link function,...). A residual R_i, regarding the i-th entry, is defined as a value that allows to quantify the

discrepancy between the observed value y_i and the fitted value $\hat{\mu_i}$. The simplest definition of residuals is the difference between the true value of the response and the fitted value given by the model ($R_i = y_i - \hat{\mu_i}$), as known as response residuals. However, as GLM apply link functions to the linear predictor, some transformations are made to the residuals with the aim of having a more correct analysis.

Pearson's residuals Pearson's residuals are the one of most simple residuals as are defined as the response residuals divided by the variance of Y_i:

$$R_i^p = \frac{y_i - \hat{\mu_i}}{\sqrt{Var(Y_i)}} = \frac{\omega_i(y_i - \hat{\mu_i})}{\sqrt{\hat{\phi} Var(\hat{\mu_i})}}. \tag{3.53}$$

Residual Deviance Another residuals are based on the deviance function:

$$R_i^d = \delta_i \sqrt{(d_i)} \tag{3.54}$$

where d_i is the parcel corresponding to the i-th observation in the deviance function and δ_i is the sign of $(y_i - \hat{\mu_i})$.
In Table 3.1 are the residuals associated with the Normal and Gamma distributions.

Table 3.1: Residuals for normal and gamma distributions

	R_p	R_d
Normal	$y_i - \hat{\mu_i}$	$y_i - \hat{\mu_i}$
Gamma	$\frac{y_i - \hat{\mu_i}}{\hat{\mu_i}}$	$\delta_i \left[2ln(\frac{\hat{\mu_i}}{y_i} + \frac{y_i - \hat{\mu_i}}{\hat{\mu_i}}) \right]^{1/2}$

Working Residuals The working residuals depend on the link function used and are defined as:

$$R_i^w = (y_i - \hat{\mu_i}) \frac{\partial \eta_i}{\partial \mu_i}. \tag{3.55}$$

The residuals analysis is very commonly made by using graphical aid. The goal is to find if the residuals show any pattern that can lead to suspicions of an eventual bad modelling decision.

- Residuals versus Predicted values
  Usually the first graph to be analysed is the residuals against the predicted values or a transformation of the predicted values. If the residuals are distributed around 0 and have similar width along the increase of the predicted values, then the model seems to be well fitted.

- Residuals versus Explanatory variable values
  Similarly, the residuals can be seen in function of the value of a certain variable. If any trend is spotted it may mean that the residuals have a dependence on the referred covariable. It can also mean that the variable used can be transformed to better suit the response.

- Residuals versus Number observation
  To check sample's independence a representation of the residuals versus the number of the observation can also be useful, and also to spot eventual outliers

- Link function
  To test if the chosen link function is well suited to the data one can create a plot of $\hat{\eta}$, the linear predictor, versus the working response u, defined as:

$$u = \hat{\eta} + \hat{D}(y - \hat{\mu})$$

  Where $\hat{D}$ is a diagonal matrix whose entries are the values of $\frac{\partial \eta_i}{\partial \mu_i}$. If the plot shows the points arranged approximately as a straight line, then the link function is well suited, if not a revaluation must be made.

3.5 Kendall's Correlation

Let (X, Y) be two joint random variables and $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ a sample from those random samples. Two pairs (x_i, y_i), (x_j, y_j) are said to be concordant if $x_i > x_j$ and $y_i > y_j$, or $x_i < x_j$ and $y_i < y_j$. If not, the pairs are said to be discordant. The Kendall's correlation coefficient τ is given by:

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \text{sign}(x_i - x_j)\text{sign}(y_i - y_j) \tag{3.56}$$

which can be interpreted as the number of concordant pairs minus the number of discordant pairs, divided by the total number of pairs $C_2^n = \frac{n(n-1)}{2}$.

3.6 Ordinal Principal Components Analysis

3.6.1 Principal Components Analysis

Principal Components Analysis (PCA) is a statistical technique used to reduce the dimensionality of the data while trying to preserve as much information as possible. In this context, preserving information can be thought as maintaining the variability associated with each entry, that is, preserving as much differences as possible. Let $X_{n\times p}$ be the data matrix, then each line can be thought of as a vector in the vectorial space $\mathbb{R}^p$. From here,

the goal is to obtain orthogonal vectors of $\mathbb{R}^p$, forming a vectorial subspace. These orthogonal vectors are what's called principal components.
Defining the first Principal Component (PC) as $\boldsymbol{a_1} = (a_{1,1},...,a_{1,p})$, then one wants to maximize $Var(\boldsymbol{a_1}'X) = \boldsymbol{a_1}'S\boldsymbol{a_1}$ where S is the covariance matrix associated with X. To make this problem solvable it is necessary to add a constraint so that the order of magnitude of $\boldsymbol{a_1}$ doesn't have any influence. Therefore, adding $\boldsymbol{a_1}'\boldsymbol{a_1} = 1$, makes $\boldsymbol{a_1}$ a unit vector, and solves the order of magnitude problem.
Thus the optimization problem can be written as:

$\max \boldsymbol{a_1}'S\boldsymbol{a_1}$
s.t. $\boldsymbol{a_1}'\boldsymbol{a_1} = 1.$

As this a non-linear constrained problem, a possible method to solve it is the Lagrangian method, and the problem becomes unrestricted, but with one more variable ($\boldsymbol{a_1}$ and λ):

$$max\, \mathcal{L}(\boldsymbol{a_1}, \lambda) = \boldsymbol{a_1}'S\boldsymbol{a_1} - \lambda(\boldsymbol{a_1}'\boldsymbol{a_1} - 1). \tag{3.57}$$

Equalling the partial derivatives to 0,

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{a_1}} = 2S\boldsymbol{a_1} - 2\lambda\boldsymbol{a_1} = 0 \tag{3.58}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \boldsymbol{a_1}'\boldsymbol{a_1} - 1 = 0 \tag{3.59}$$

leads to:

$$\begin{cases} \boldsymbol{a_1}'\boldsymbol{a_1} = 1 \\ S\boldsymbol{a_1} = \lambda\boldsymbol{a_1} \end{cases}. \tag{3.60}$$

One can finally conclude that the λ that maximizes the objective function and respects the conditions is the first eigenvalue of S and $\boldsymbol{a_1}$ is the respective eigenvector.The first principal component is, then, the eigenvector associated to the largest eigenvalue, the second PC is the eigenvector of the second largest eigenvalue and the same logic applies until the last eigenvalue.
As the main goal is to reduce the dimensionality of the data, it is not intended to use all the PC's, as in this case the dimensionality would be the same. A helpful metric to choose how many PC's to use it to evaluate the percentage of variance that each PC retains. This can be achieved by dividing the respective eigenvalues by p (as the sum of all eigenvalues is the order of the matrix S). One criterion is to choose the number of PC's that allow to explain from 70 to 90% of all variance. The Kaiser criterion states that one must choose the number of PC's that allow to explain between 70 to 90% of the variance. One other possible method is to use the Screeplot, a plot with the PC number on the horizontal axis and the eigenvalues on the vertical one. The ideal number of PC's is chosen by stopping

where the plot line becomes, approximately, horizontal.
There are, however, some considerations that must be taken into account when using this technique. The first is that it is based on the existence of high correlations between the variables, if the variables are approximately uncorrelated, then, PCA will not be very useful. One way to check this is by using Mauchly's test of sphericity, where the correlation matrix is compared with the identity matrix, which symbolizes uncorrelated variables. Other important reflection is the order of magnitude of the variables. If the variables have very different order of magnitudes, then, standardizing them is a very pertinent option. Finally, it is important to note that this technique has better results when all variables are continuous, if not, some adaptations can be made to improve the results.

3.6.2 Categorical PCA

As previously mentioned, PCA applied to nominal or ordinal data has its own challenges as standard PCA may return poor results. In both cases the common technique to represent each category is by assigning a number to it. These values may have some interpretation, in the ordinal case, or not, as in nominal variables. However, in both cases, the chosen scale might not be the one that best fits in a certain model. To solve this problem Optimal Scaling is an important tool to increase the relation between the scales and the model. The main principle of optimal scaling is simply to apply a function to the original scale that minimizes the difference between each category's observations and the overall behavior of those observations. These functions can be of three types:

- Nominal transformation - This transformation is the one that allows more freedom, as there are no constraints attached. Each original value can change to any other value.
- Ordinal transformation - Ordinal transformations follow the same principle but are restricted to transformations that do not change the original order of each category.
- Metric transformation - This last transformations are the ones with more restrictions, as besides the order of the categories, also the distance between categories must be the same. In resume, a linear transformation is applied to the original scale.

The last transformation can also be extended to other non linear functions, like quadratic or cubic, but the most useful are monotone splines, as allow more control of each category's score.

3.6.2.1 Gifi methods

Gifi methods are optimal scaling methods that focus on dimensionality reduction [34, 35], just like classic PCA. Therefore, the initial step is to consider the data already with

dimensionality reduction, with p dimensions. Let $H_{n\times m}$ be the original data matrix, h_j the column vector associated with each variable and k_j the number of different categories in each variable, where $j \in \{1,...,m\}$. Let G_j be the indicator matrix, $n \times k_j$, consisting of 0's and 1's, each line has a 1 in the column correspondent to the entry's category value and 0 in all others, again with $j \in \{1,...,m\}$, one can also define $D_j = G_j' G_j$. For each variable it is defined Y_j, $k_j \times p$, be the category quantifications matrix, that gives each category its own "point"in the p-dimensional space. Finally, let $X_{n\times p}$, be the dimensionally reduced matrix, called objects scores matrix, where are each entry's coordinates in the p-dimensional space.

The first step of the procedure is to calculate Y_j, the quantifications matrix. For each $j \in \{1,...,m\}$ and $i \in \{1,...,k_j\}$, the i-th line of Y_j is the centroid of all the points who belong to the i-th category of variable j. From here, the objective function can be written as:

$$\sigma(X, Y_1, ..., Y_m) = \sum_{j=1}^{m} tr[(X - G_j Y_j)'(X - G_j Y_j)].$$

The goal is to minimize the objective function, which translates to minimizing the sum of squares of the distance between each category's centroid and the points of that category. In addition two other constraints must be considered:

$$X'X = NI_p$$
$$u_N' X = 0.$$

The first constraint standardizes the squared length of the object scores and the columns of X to be orthogonal, a very important trait in PCA. The second one, centers the plot in the origin, where u_N' is a vector of ones.

It is important to note that this optimization process is made in two directions, first the calculation of the centroids is made, after that, the values of X are changed to be closer to the centroids, by making the score of an entry equal to the mean of the centroids of the categories it belongs to. However, as these values are changed, the centroids will change, and therefore the X values have to be updated, and so on. To solve this problem an Alternating Least Squares (ALS) algorithm is used:

This solution is called Homogeneity Analysis by Means of Alternating Least Squares (HOMALS) solution, and is used with nominal data.

Algorithm 3 Alternating Least Squares - HOMALS

while Stop condition == FALSE **do**
- $\hat{Y}_j = (D_j)^{-1} G_j' X, \forall j \in \{1,...,m\}$ ▷ Calculate category centroids
- $\hat{X} = \frac{1}{m} \sum_{j=1}^{m} G_j \hat{Y}_j$ ▷ Update X's values
- $W = \hat{X} - u_N (u_N' \hat{X}/N)$ ▷ Center scores
- $X = \sqrt{N} GRAM(W)$ ▷ Apply Gram-Schmidt algorithm

end while

This solution is very helpful in the case of nominal data, but gives no guarantees about ordinal or numeric data, as there's no restriction of these type, and therefore, in those cases, an additional constraint has to be introduced:

$$Y_j = q_j \beta_j'.$$

Where q_j, $k_j \times 1$, is a column matrix with single category quantifications for variable j, and β_j, $p \times 1$, is a vector of weights (component loadings). This restriction will guarantee that Y_j is of rank 1, meaning that each row is a linear combination of the first (and of all rows, consequently), which assures that the order or the distance between the centroids of each category are maintained. As, again, q_j and β_j depend on each other, another ALS cycle is added. The update algorithm is as follows:

This solution is known as the Principal Components Analysis by means of Alternating Least Squares (PRINCALS) solution and is applied to ordinal or numeric data.

Algorithm 4 Alternating Least Squares - PRINCALS

while Stop condition == FALSE **do**

$\hat{Y}_j = (D_j)^{-1} G_j' X, \forall j \in \{1, ..., m\}$ ▹ Calculate category centroids

$\hat{\beta}_j = (\hat{Y}_j{}' D_j q_j)/(q_j' D_j q_j)$ ▹ Estimate Component Loadings

$\hat{q_j} = (\hat{Y}_j \beta_j)/(\beta_j' \beta_j)$ ▹ Estimate simple category quantifications

Monotone(or linear) regression ▹ Guarantee that $\hat{q_j}$ respects the order (or distance)

$\hat{Y}_j = \hat{q_j} \hat{\beta}_j{}'$ ▹ Update category centroids

$\hat{X} = \frac{1}{m} \sum_{j=1}^{m} G_j \hat{Y}_j$ ▹ Update X's values

$W = \hat{X} - u_N (u_N' \hat{X}/N)$ ▹ Center scores

$X = \sqrt{N} GRAM(W)$ ▹ Apply Gram-Schmidt algorithm

end while

3.7 Prediction Quality

Many models have as main goal to predict outcomes, thus there are methods that quantify the quality of the predictions made by the model. If possible, the sample should be different then the one used to develop the model, but, as in many cases obtaining quality data is not easy, the application may be made on the same sample.

3.7.1 Receiver Operating Characteristic curve

Receiver Operating Characteristic (ROC) curve is a common tool used to measure the quality of a binary classifier. This section is based on Gonçalves et al. [36].
Let X and Y be two independent random variables representing the score of a medical test for the healthy population (D=0) and for the population suffering from the disease (D=1), respectively. Let c be the cut-off value for which the test result is positive if the score is greater than c and negative otherwise. Let F and G denote the probability density functions of X and Y respectively. The sensitivity of the test is given by $Se(c) = 1 - G(c)$, and the specificity by $Sp(c) = F(c)$. The ROC curve is a plot of $Se(c)$ versus $1 - Sp(c)$, for $-\infty \leq c \leq \infty$, which is the equivalent to:

$$ROC(t) = 1 - G(F^{-1}(1-t)),\ t \in [0,1] \tag{3.61}$$

where $F^{-1}(1-t) = \inf\{x \in \mathbb{R} : F(x) \geq 1-t\}$. An important metric associated with ROC curves is the area under the curve (AUC) which is defined as:

$$AUC = \int_0^1 ROC(u)du \tag{3.62}$$

and can be interpreted as the probability of, given one healthy person and one diseased person (chose randomly), the diagnostic test value is higher for the diseased individual (i.e. $P(Y > X)$).The ideal classifier would be one with 100% of sensitivity and 100% of specificity, which would mean an AUC of 1, as $P(Y > X) = 1$. On the other hand, a random classifier would assign random values for each individual which in the population would lead to an AUC of 0.5, as $P(Y > X) = 0.5$
In this particular thesis it is used the non-parametric estimation of the ROC curve that uses the empirical distributions and quantile functions associated with the healthy and diseased populations.

$$\tilde{ROC}(t) = 1 - \tilde{G}(\tilde{F}^{-1}(1-t)),\ t \in [0,1] \tag{3.63}$$

where $\tilde{F}^{-1}$ represents the empirical quantile function of healthy population and $\tilde{G}$ the empirical distribution function of the diseased population. The empirical distribution function is, approximately, defined, for any given value t, as the percentage of sample points smaller or equal to t.
Given a fix threshold c, a confusion matrix can be constructed, which relates the outcome

Table 3.2: Confusion matrix of a binary classifier

		Predicted Outcome	
		Positive (PP)	Negative (PN)
Real Outcome	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

of the test (predicted) with the real outcome. From the confusion matrix is possible to retrieve the following information helpful for assessing the model and threshold performance.

- Sensitivity or True positive rate (TPR) - $\frac{TP}{P}$
- Specifity or True negative rate (TNR) - $\frac{TN}{N}$
- Precision or Positive predicted value (PPV) - $\frac{TP}{PP}$
- Accuracy - $\frac{TP+TN}{P+N}$

The definition of the threshold is a big challenge because it influences sensitivity and specificity in an opposing way. Regarding AUC the value is given without the need for a fixed threshold and can compare the performance of the classifier with the ideal situation (AUC=1) and the performance of the random classifier (AUC=0.5).

3.8 Software

The software used to develop this study was R [37], trough RStudio [38]. The ROC curves construction was made with the help of packages ROCR [39] and pROC [40], while the plots were made using the packages ggplot2 [41], RColorBrewer [42], ggpubr [43] and qqplotr [44]. The thesis template used was the novathesis Latex template [**novathesis-manual**].

4

Application

4.1 Data Description

The lung ultrasounds made available are divided in 4 zones in each lung and each zone has 4 scores associated with it. B-lines are scored 0, 1, 2 or 3 from better to worse. Pleura, Subpleural Consolidations (SubP) and Lobar Consolidations (Lob) are scored with 0 or 1, 1 representing heterogeneous pleura, existence of subpleural consolidation and existence of lobar consolidation, respectively.

4.1.1 Frequencies

Figure 4.1 shows the distribution of the B-lines scores by lung zone. Figure 4.2 depicts the distribution of the totals, that is, the total of the right lung (T_R), total of left lung (T_L) and the overall total. The plots show that the middle scores are the most common, around at least 70% in all zones, and the most severe score is the less used. In terms of differences between zones the most obvious is that the right lung has scores with slightly less severity, as the mode is the score 1 when compared to the score 2 of the left lung.
A similar analysis for Pleura, SubP and Lob was made with the results shown in figures 4.3, 4.4, 4.5, respectively.
The existence of heterogeneous pleura appears to be more frequent in zones 3 and 4 of each lung, over 92% of cases in these zones, when compared to zones 1 and 2 where the incidence is always around but below 90%. These values show that heterogeneous pleura is very common in hospitalized patients, which may indicate that it won't be a very good indicator to predict severity. In regard to differences between lungs, there is no evidence that they follow different distributions.
Concerning subpleural consolidations the same zone has similar results in each lung with the exception of zone 4 where there's a substantial difference between both lungs. There are also significant differences between zones 1 and 2, and zones 3 and 4, with the latter having higher incidence when compared to the former. It can also be seen that in every zone the left lung has a higher incidence of these consolidations, which is corroborated by the lung comparison, where the right lung has an average of 20% while the left has an

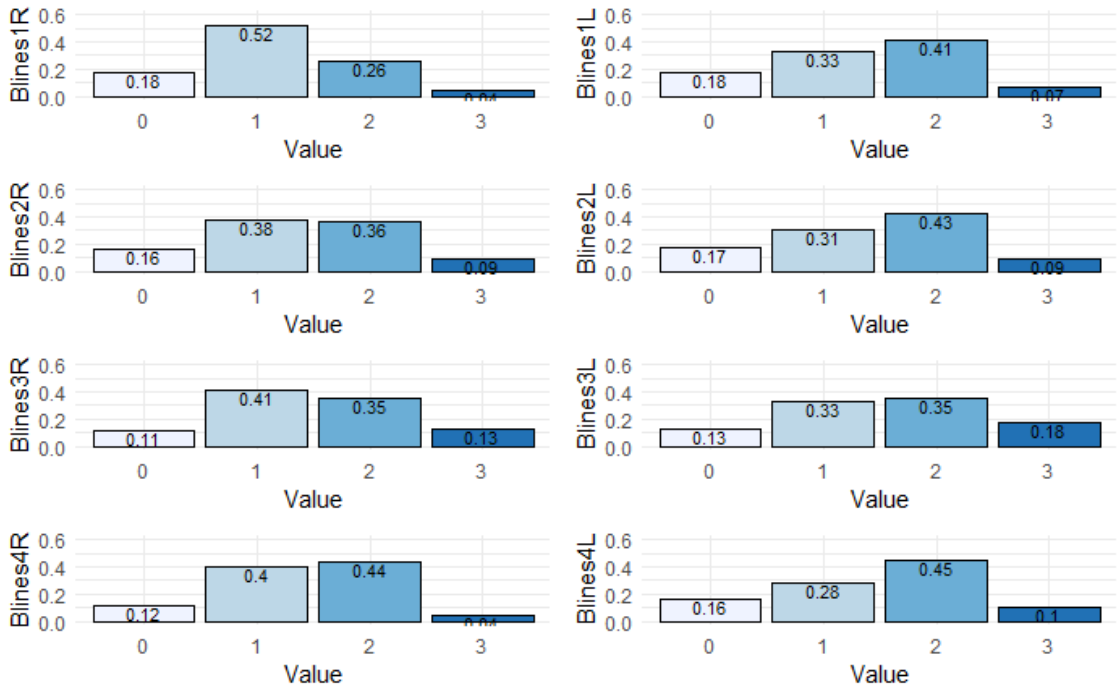

Figure 4.1: Distribution of the B-lines scores distribution by lung zone

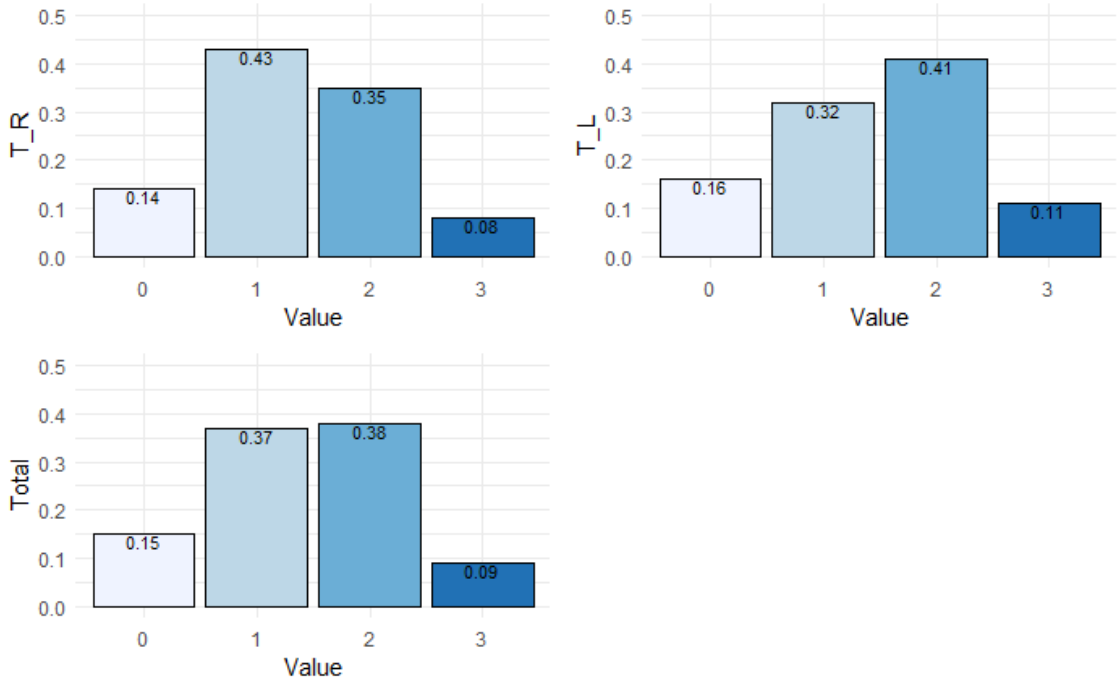

Figure 4.2: Distribution of the B-lines scores by lung (right, T_R, and left, T_L) and in total (Total)

average of 27%, regarding SubP consolidations.

Finally, lobar consolidation is the rarest of these 4 variables, with an overall prevalence of

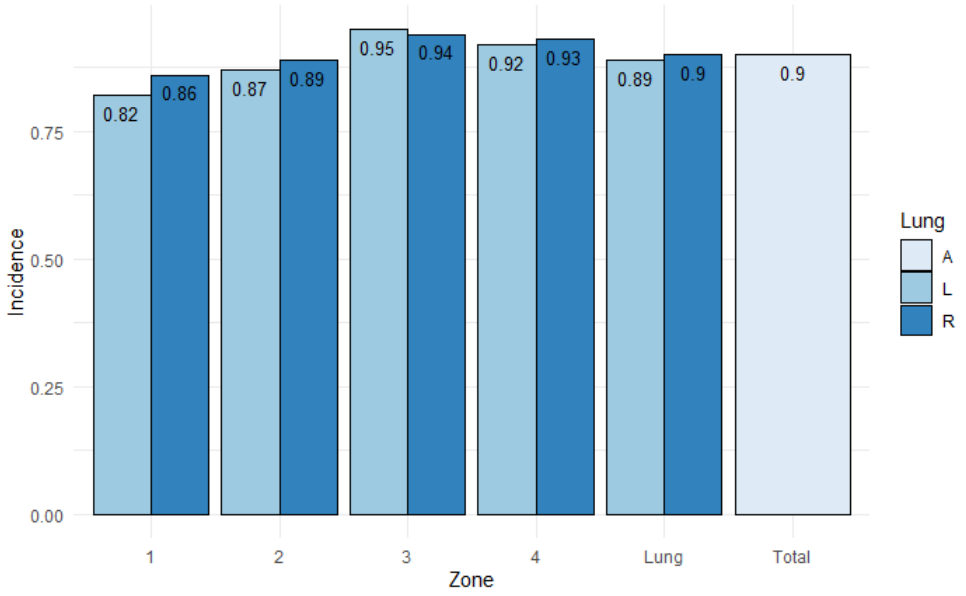

Figure 4.3: Incidence, in proportion, of heterogeneous Pleura by lung zone (1,2,3 and 4), by lung (L,R) and in Total(A)

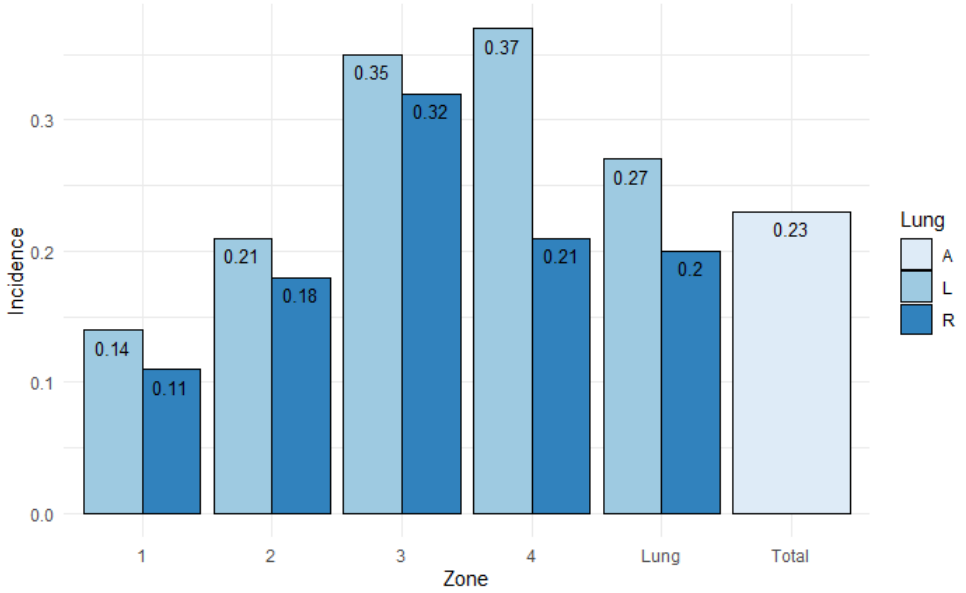

Figure 4.4: Incidence, in proportion, of SubP by lung zone (1,2,3 and 4), by lung (L,R) and in Total(A)

only 3%, and zone 4R has not registered a single case. Nonetheless, there may be sufficient evidence that these consolidations affect more frequently the left lung, 5%, than the right one, 1%.

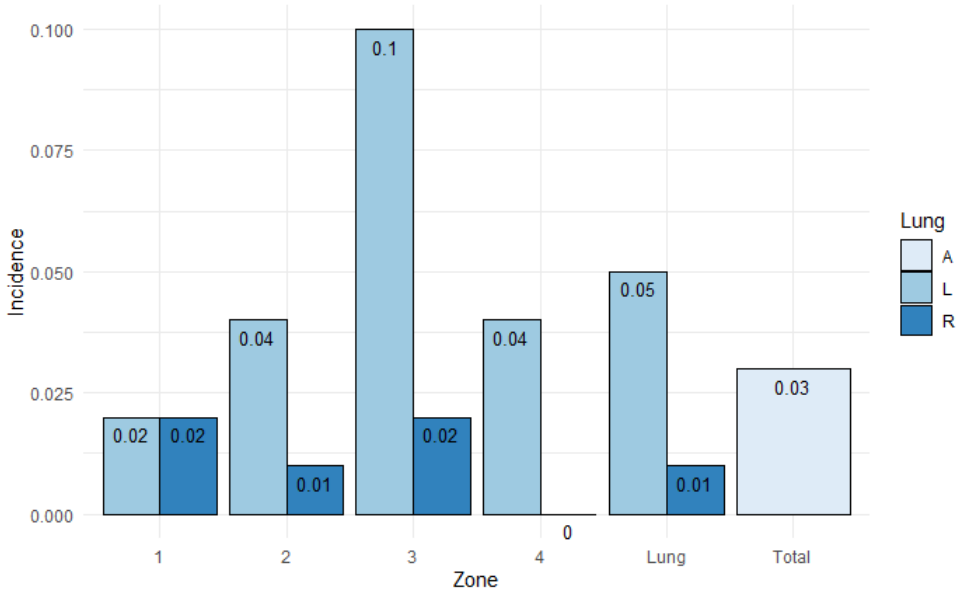

Figure 4.5: Incidence, in proportion, of Lob by lung zone (1,2,3 and 4), by lung (L,R) and in Total(A)

4.1.2 Length of Stay and Intensive Care Unit Length of Stay

As referred in chapter 2, Length of Stay (LOS) is a very important variable when modelling a disease's severity. Therefore, an analysis of the overall LOS, of the LOS of the patients that didn't die and of the Intensive Care Unit (ICU) LOS is made, along with the best fitted theoretical distributions.
In figure 4.6, one can see the histogram related to the LOS of all patients, along with the more appropriate approximate distributions and the respective p-values of the goodness of fit test, Kolmogorov-Smirnov (KS) for Gamma and Shapiro-Wilks for Log-normal (the Shapiro-Wilks test was applied to the logarithm of the values).
Figure 4.6 shows that, the majority of the sample's entries have small values of LOS, which leads to a right-skewed distribution.Two known right-skewed distributions are the Log-normal and the Gamma distributions. In this particular case, the Log-normal distribution has a better goodness of fit p-value compared to the Gamma distribution (0.72 vs 0.41), but both will be considered as the distribution of the response variable in Generalized Linear Model (GLM)'s, as either one can be model by this technique. These p-values are very encouraging indicators for the performance of GLM's, because the better the fit, the more adequate is the application of this methodology.
In figure 4.7 is the distribution of LOS of the patients that left the hospital alive. As the sample has very few deaths, there is little difference between the two histograms, and, again, both distributions have a reasonable good fit to the data.
Finally, figure 4.8 shows the distribution of the ICU LOS for the patients that actually went to ICU. Similarly to the other two, the data is right-skewed but in this case both

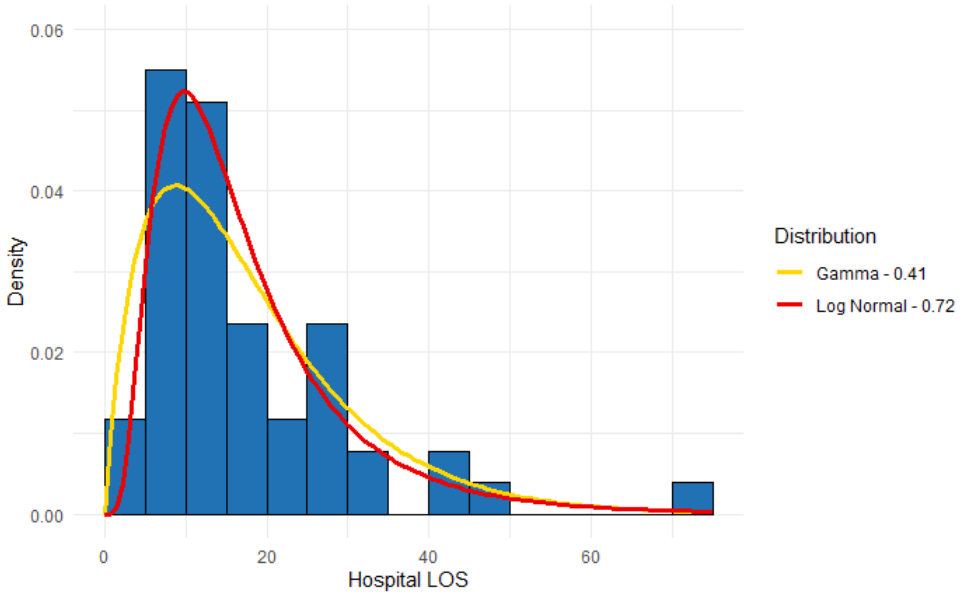

Figure 4.6: LOS empirical distribution (blue histogram) and estimated Gamma (yellow curve) and Log-normal (red curve) parent distributions, with the respective p-values for the KS test

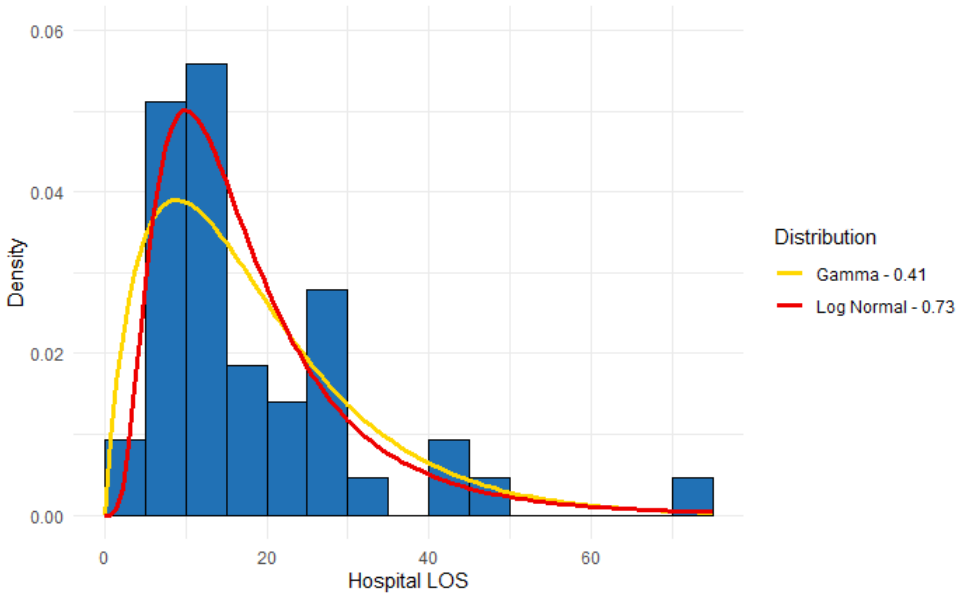

Figure 4.7: LOS of alive patients empirical distribution (blue histogram) and estimated Gamma (yellow curve) and Log-normal (red curve) parent distributions, with the respective p-values for the KS test

theoretical distributions have almost equal p-values (0.46 for Gamma and 0.45 for Log-normal).

Figure 4.9 presents the box plots associated with these three variables, where it shows that the alive patients tend to be hospitalized slightly more time (18.7 days vs 18.0, on average) and with greater variation as the standard variation is 12.9 for the general patient and 13.5 for the alive ones. The ICU LOS is smaller because the time of ICU hospitalization

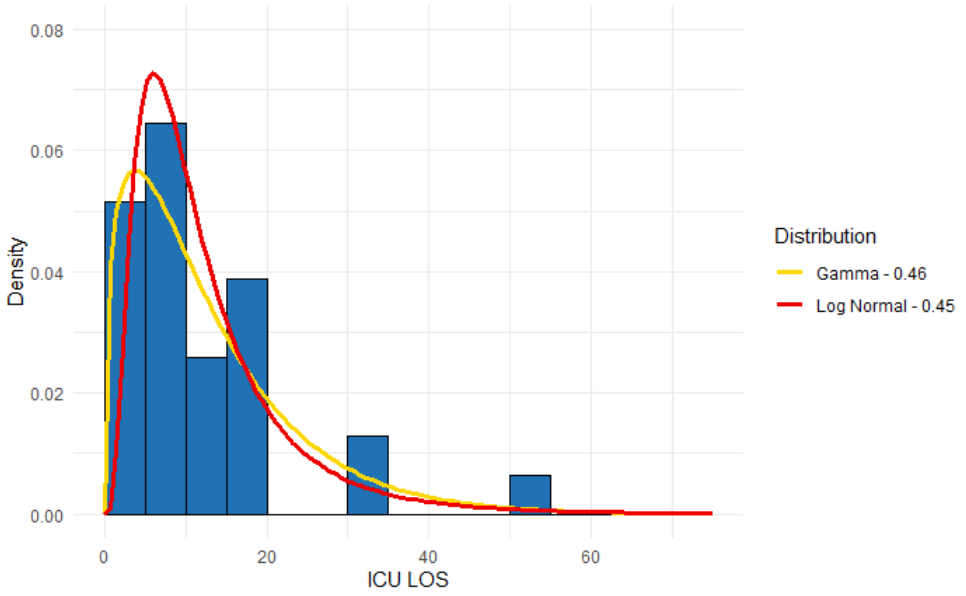

Figure 4.8: ICU LOS of ICU hospitalized patients empirical distribution (blue histogram) and estimated Gamma (yellow curve) and Log-normal (red curve) parent distributions, with the respective p-values for the KS test

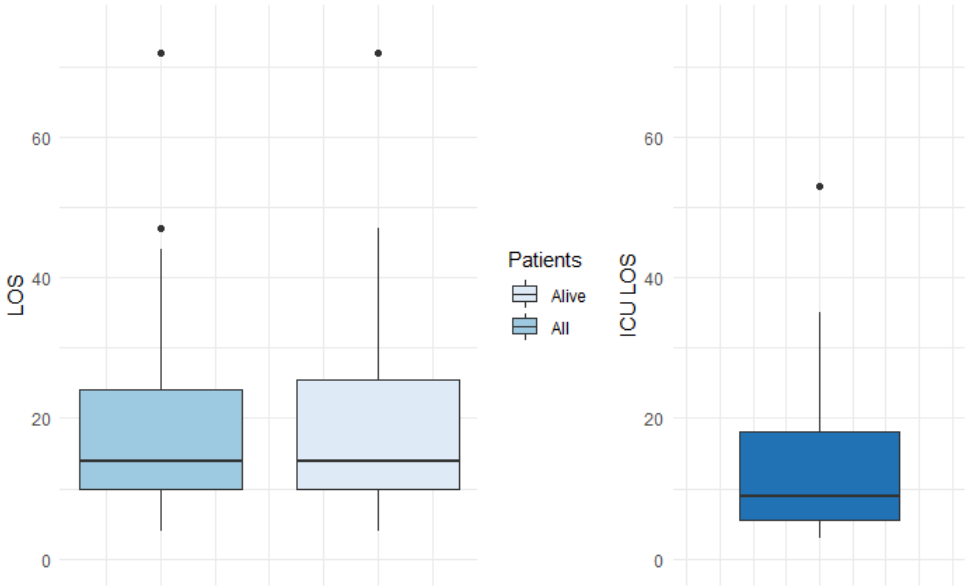

Figure 4.9: LOS of all patients (blue box plot), LOS of alive patients (light blue box plot) and ICU LOS of ICU hospitalized patients (dark blue box plot) box plots

is included in the overall LOS and also because hospitalization in this unit is only made when the patient is in severe condition.

In conclusion, both Log-normal and Gamma distributions present a good fit to the LOS observed distribution, in general and in ICU. Apart from a possible outlier in the right tail, the histogram values and the curves are closely related. In overall LOS the Log-normal distribution appears to have a better fit when compared to the Gamma distribution, but

in ICU LOS they have similar results, as previously mentioned.

4.1.3 Mortality, Length of Stay and ICU Length of Stay by variable and Lung Zone

To quantify the severity of a disease in a given patient several variables can be acknowledge. Usually, the first studied outcome is whether the patient died or not, which is often represent by a binary variable (1 meaning death and 0 meaning the patient lived). The patients can, however, experience different symptoms and with different levels of severity. A possible variable to differentiate these patients is the hospital LOS, which is based on the idea that the longer a patient stays in the hospital the more severe is the effect of the disease. In a similar chain of thought one can affirm that, in general, the patients that spend more time in ICU have more severe cases when compared to those that stay in the nursery the whole time.

4.1.3.1 Mortality

Figure 4.10 shows the comparison of the mortality rates between the B-lines scores in each lung zone, with the respective 95% percentile confidence interval, built by 1000 bootstrap simulations. It can be seen that, in general, the increase of the score also leads to an increase in the mortality. However, especially in the left lung, the score 0 has higher mortality rate than 1. This may be explained by the fact that the first ultrasounds can have a little affected lung (with some zones scored 0) but the worsening of the patient leads to its death, and therefore an ultrasound with zeros is associated with a patient that died. >The relatively small sample can also be an explanation, as can be seen by the large confidence interval that in almost all cases has a lower bound of zero.

A similar analysis can be made for the binary variables. In figure 4.11, the mortality according to the pleura evaluation is presented. As was shown in the previous section, the majority of the ultrasounds showed heterogeneous pleura and, therefore, the mortality rates for the score 1 are more consistent, as can be seen by the smaller confidence interval. The rates for this score are similar in all lung zones, about 10%, and the major difference appears in the cases of homogeneous pleura (score 0). The rates differ a lot from zone to zone and the confidence intervals are very long, therefore no strong conclusions can be taken. Even so, in 5 zones the mortality is smaller when compared to the heterogeneous cases and there is just one zone where the mortality is clearly higher for the homogeneous pleura (1R).

Concerning the SubPleural Consolidations, figure 4.12 shows that in most zones the mortality is higher when these consolidations are present. In the zones where the mortality is higher when the consolidations are absent the confidence interval shows that the difference between the absence and the presence is within the margin of error. Finally, the lobar consolidations are the ones that have higher mortality of all the scores analysed

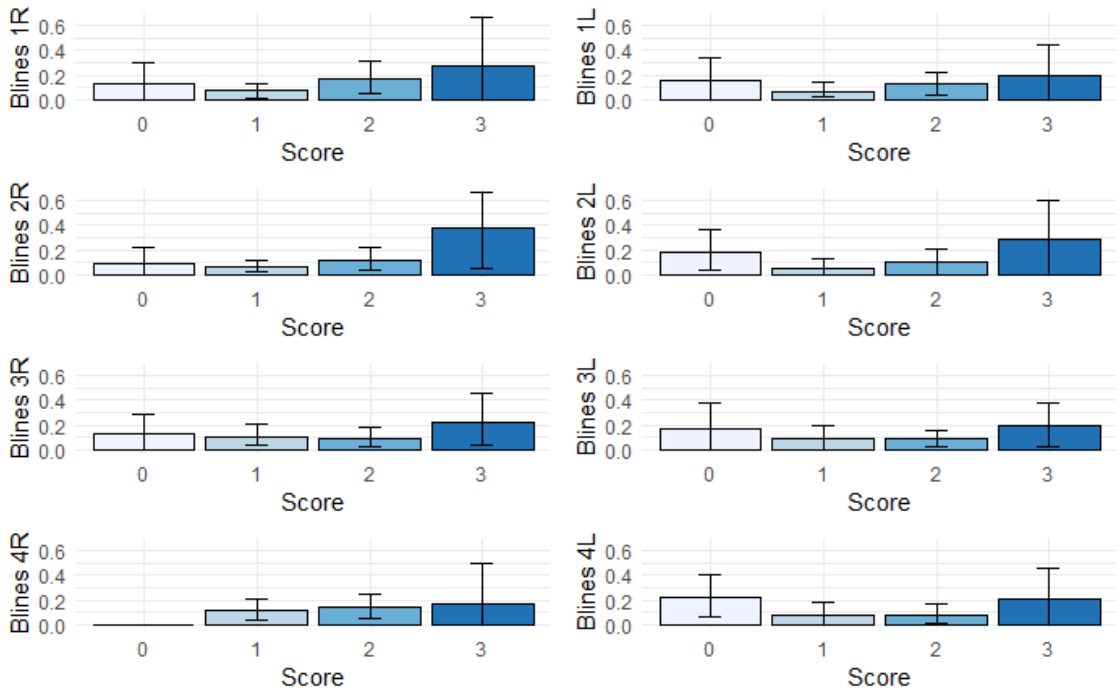

Figure 4.10: Mortality proportion by B-lines' score and lung zone with respective confidence intervals obtained by bootstrap

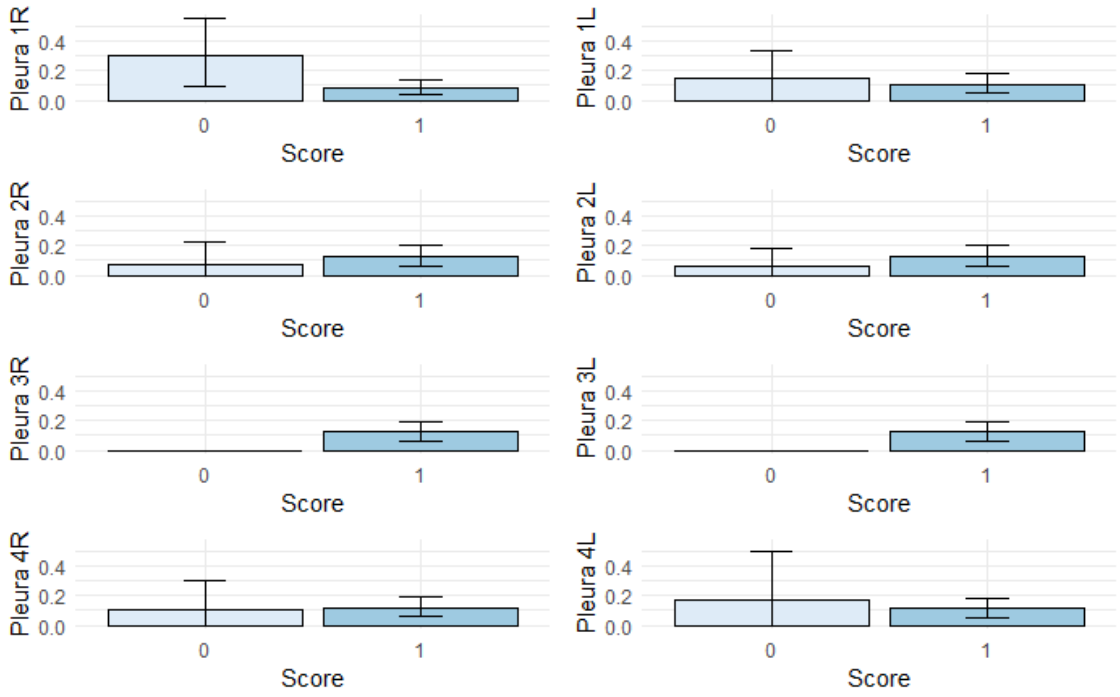

Figure 4.11: Mortality proportion by Pleura's score and lung zone with respective confidence intervals obtained by bootstrap

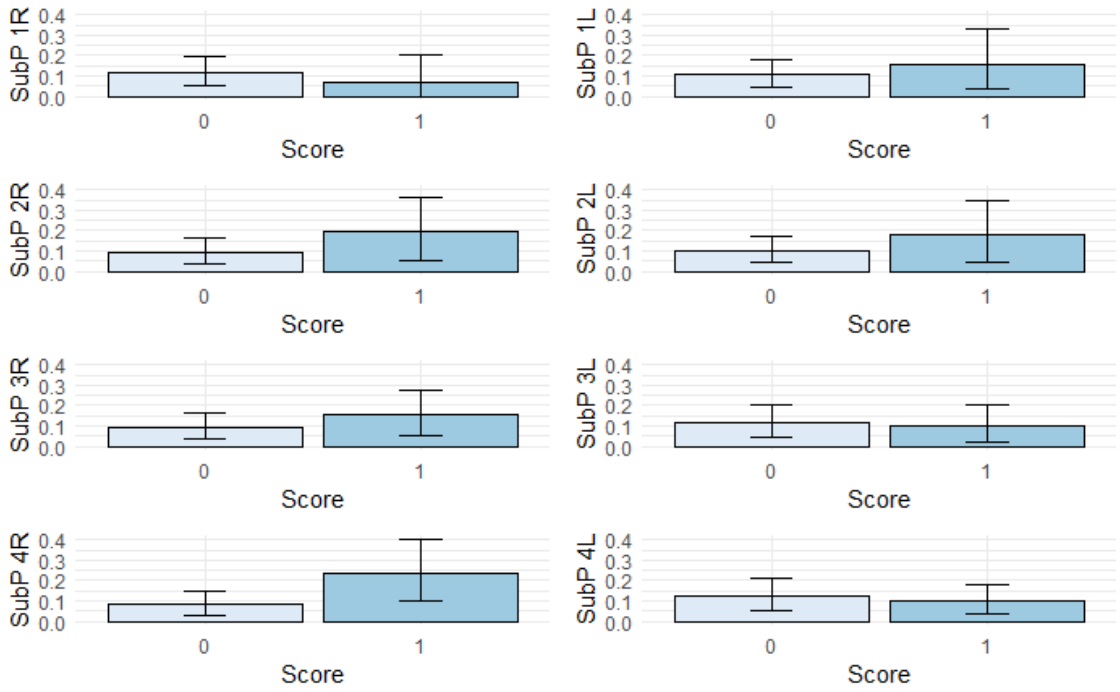

Figure 4.12: Mortality proportion by SubP's score and lung zone with respective confidence intervals obtained by bootstrap

(figure 4.13). As these consolidations are relatively rare, about only 3% of incidence, the corresponding mortality rates have a large confidence interval. Nevertheless, the mortality when consolidations are present is clearly higher when compared to the cases where these are absent. It also appears that the ones in the left lung have more influence on mortality, however, this may be caused by the small sample, since the incidence on the right lung is only of 1%.

In conclusion, in most cases the increase of the score leads to an increase in the mortality rate, which means that these variables might provide valuable information to predict the patients' outcomes.

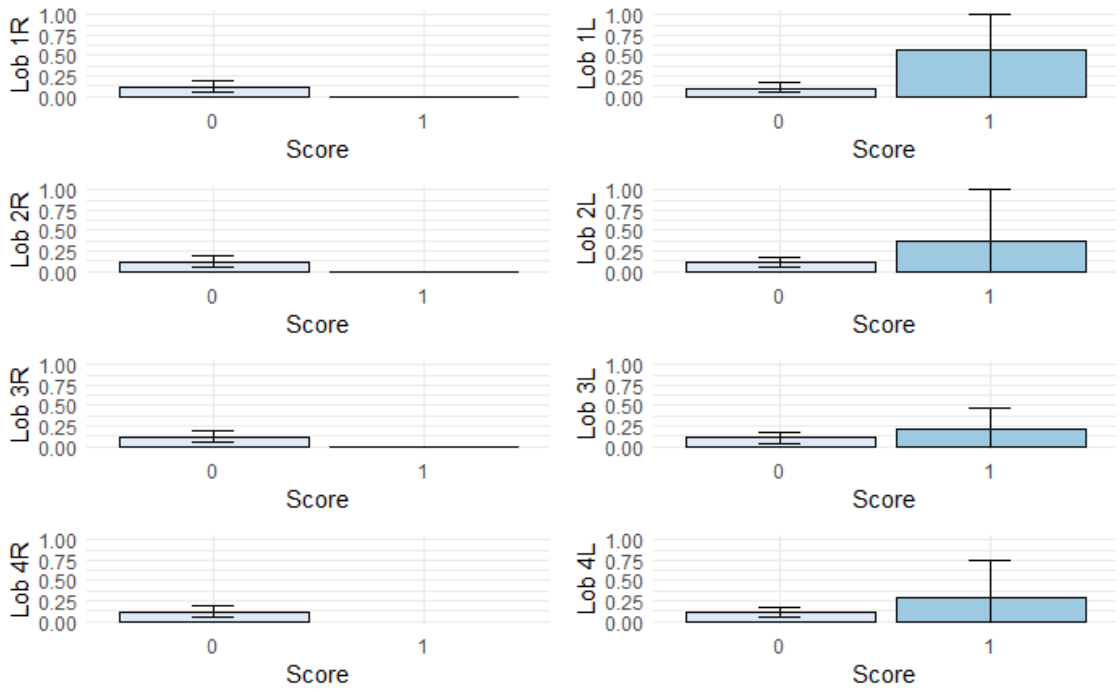

Figure 4.13: Mortality proportion by Lob's score and lung zone with respective confidence intervals obtained by bootstrap

4.1.3.2 Length of Stay

The hospital LOS is an important variable to separate severe from mild or moderate cases as the severe ones tend to have longer hospitalizations. It is important to note, however, that LOS is not necessarily a good differentiator between death and non-death. For example, a patient may die in a small number of days while other patient may live after a long hospitalization. In this particular example, the patient that passed away has a smaller LOS although he had a more severe case when compared to the patient that lived.

The differences in the average LOS by B-lines' scores are a perfect example of what was explained in the previous paragraph (figure 4.14). The LOS increases from 0 to 1 and from 1 to 2 in all zones (except 4R and 1L, where 1 and 2 have approximately the same average), which shows that the severity increase leads to longer hospitalizations. From 2 to 3, though, there is a decrease in the LOS which is explained not because the severity is smaller but, perhaps, because the increase of the mortality leads to smaller hospitalization time. The effect of Pleura's scores has less influence in the LOS, figure 4.15, as the differences between absence and presence of heterogeneous pleura are small and within the confidence intervals. In general, however, it can be said that the presence leads to slightly higher hospitalization time. Again, it is important to note that the class of heterogeneous pleura represent 90% of the sample, and this unbalance leads to high confidence intervals in the class of 0.

Regarding the Subpleural consolidations, there are more consistent results as their pres-

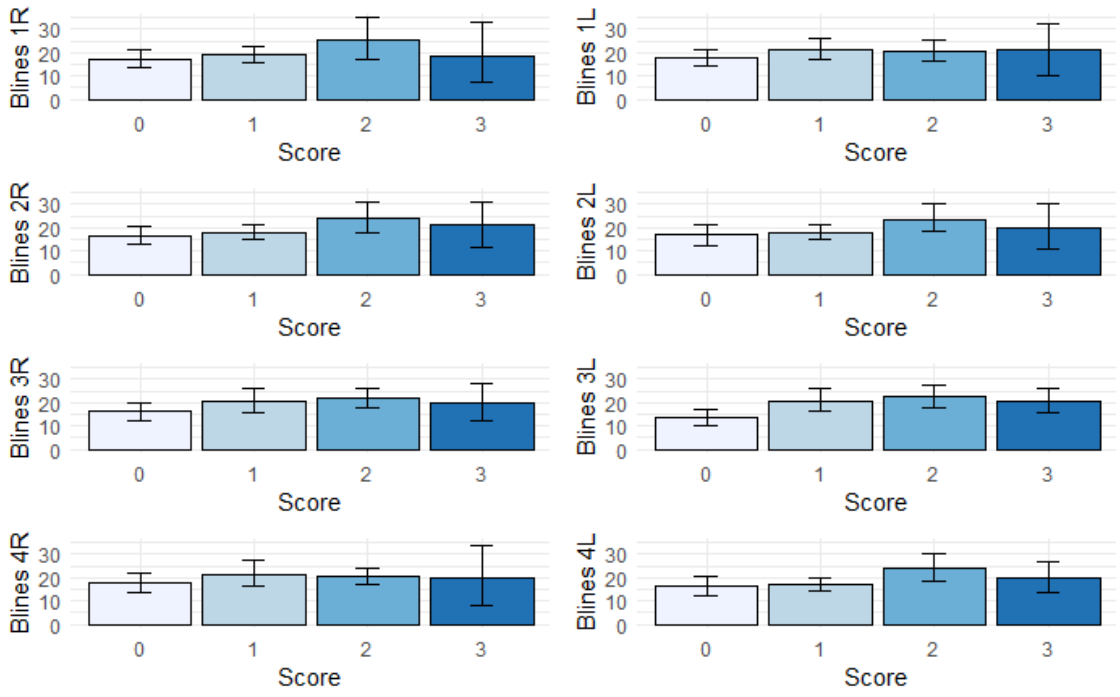

Figure 4.14: Average LOS for each B-lines' score and lung zone with respective confidence intervals obtained by bootstrap

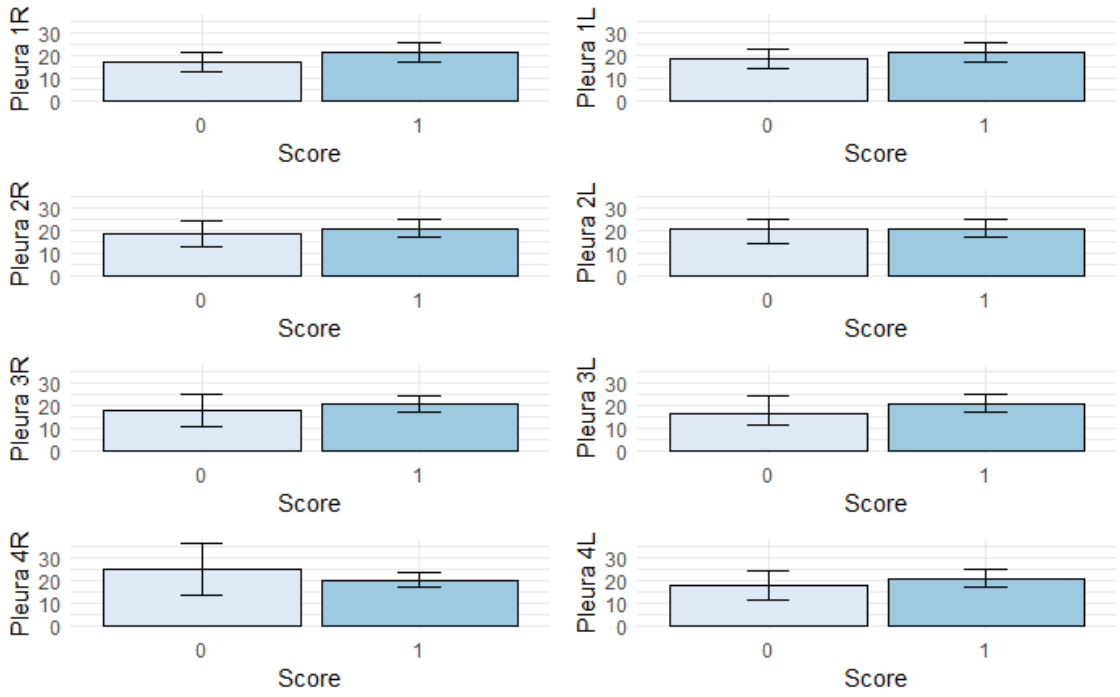

Figure 4.15: Average LOS for each Pleura's score and lung zone with respective confidence intervals obtained by bootstrap

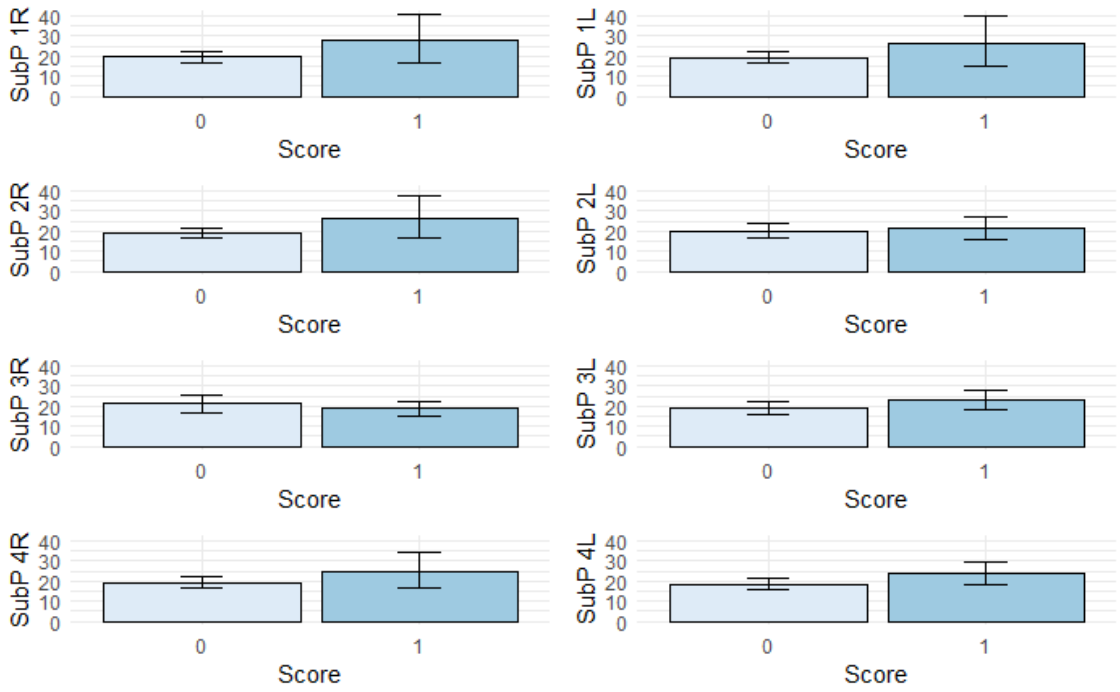

Figure 4.16: Average LOS for each SubP's score and lung zone with respective confidence intervals obtained by bootstrap

ence leads to a significantly higher LOS, as in most zones the average of the score 1 is bigger than the upper bound of the confidence interval regarding the class of 0, as seen in figure 4.16. As the mortality rates related to the presence of these consolidations are not that high, the increased severity on patients that do not die has a bigger influence than mortality on the LOS, which explains the increase.

Again, the lobar consolidations have unstable results given their small incidence, figure 4.17. In all zones the 0 value has a similar LOS average and a very small confidence interval. On the other hand, in the presence of lobar consolidations, each zone has very different results, from zone 4R where no consolidations were recorded to zone 1R where the mean is almost 50 days and the upper bound reaching 70 days, with the length of the confidence interval being of 60 days.

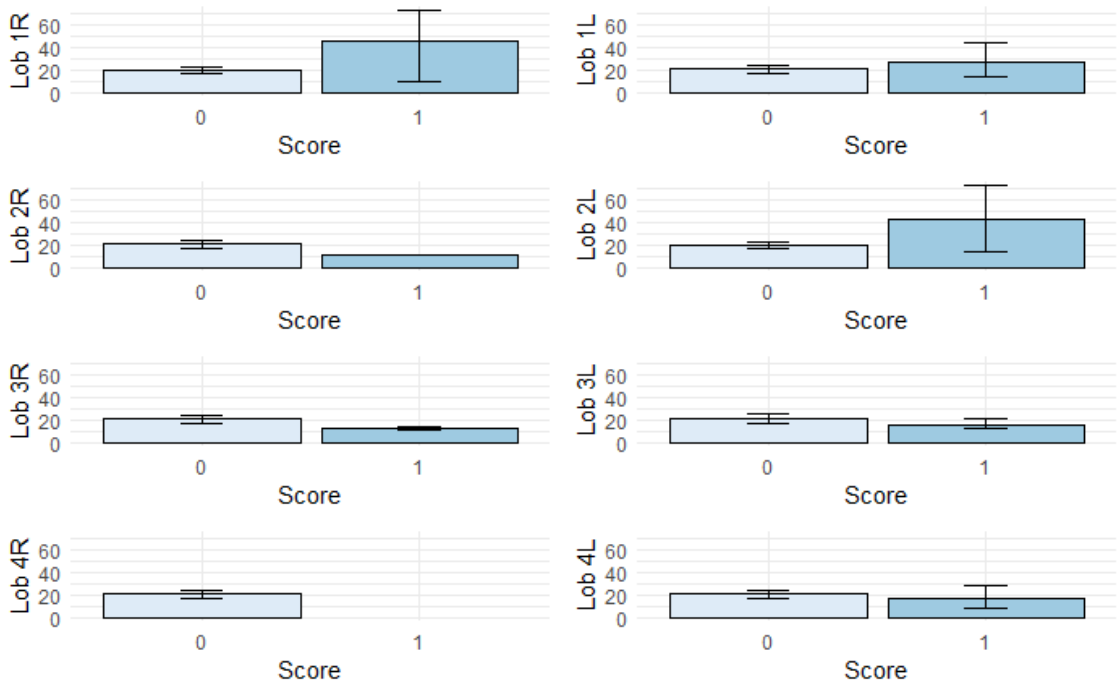

Figure 4.17: Average LOS for each Lob's score and lung zone with respective confidence intervals obtained by bootstrap

4.1.3.3 ICU Length of Stay

In a similar way, ICU LOS can give valuable information about severity, but caution must be taken given the influence death has on LOS and also that some patients characteristics do not allow them to be in a ICU, for example if the patient is already in terminal stage and would not benefit of ICU hospitalization. The B-lines' scores influence on ICU LOS has some resemblances with their influence on overall LOS, figure 4.18. The increase from 0 to 1 and 1 to 2 is similar, but in this case the relation between score 2 and score 3 is a slightly different. In this case, there are five zones where average ICU LOS is bigger in score 2 and three zones where is bigger in score 3. Both are always higher than ICU LOS of score 1, which didn't happen in LOS where there were cases where score 1 and score 3 had similar values. This leads to the conclusion that this variable is even more influenced by severity and the patient's outcome (death or non-death) has slightly less impact.

In the Pleura's variable, the same conclusion can be taken from overall LOS. The differences between the scores are small, in 5 zones score 1 has higher average and in the vast majority of zones the confidence intervals have a large proportion of overlapping values, as can be seen in figure 4.19.

Figure 4.20 highlights the influence that the existence of Subpleural consolidations have on the average ICU LOS. In this case, the effect is clearer than in the overall LOS, as in all zones the average is higher when the score is 1 and the confidence intervals are very small in score 0, which leads to the conclusion that these consolidations increase significantly

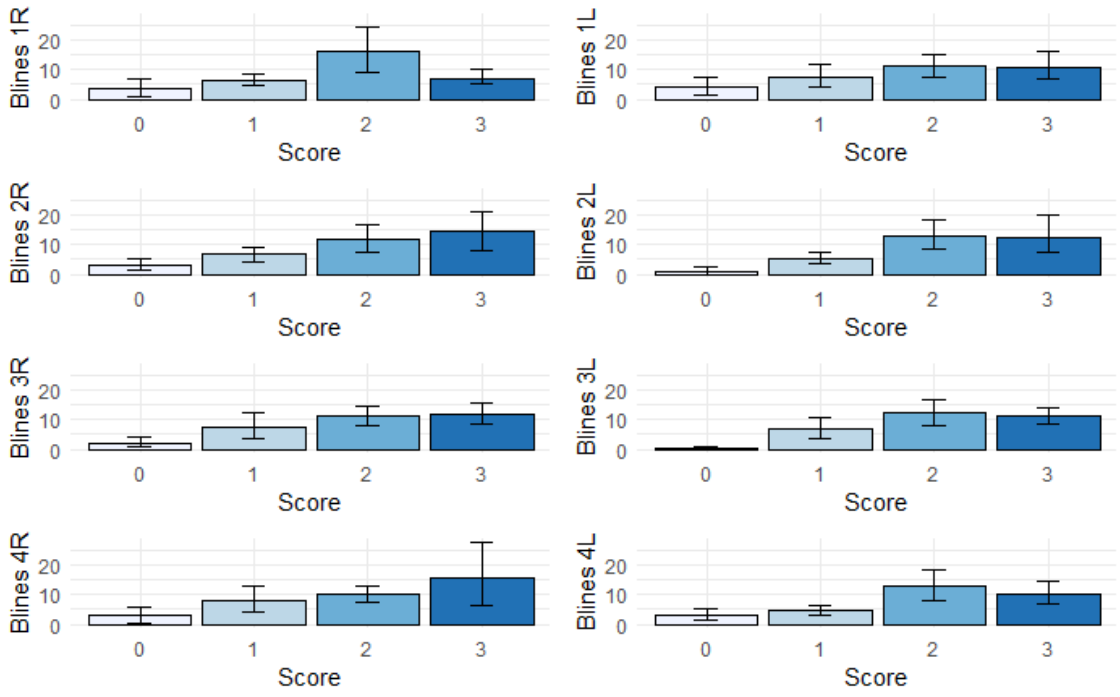

Figure 4.18: Average ICU LOS for each B-lines' score and lung zone with respective confidence intervals obtained by bootstrap

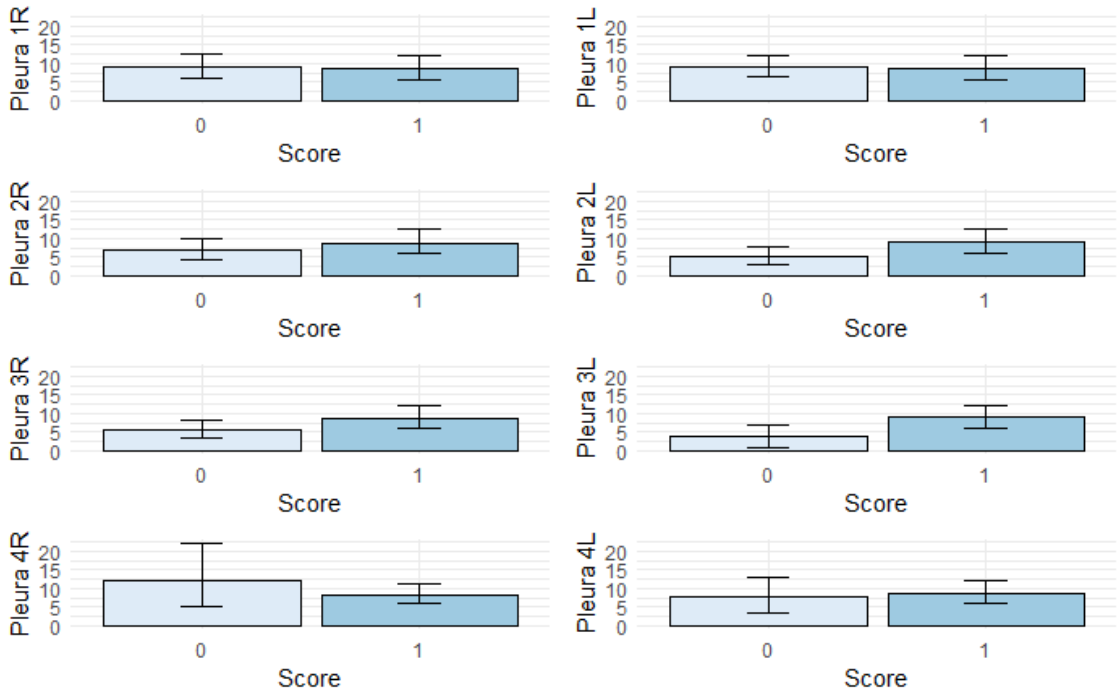

Figure 4.19: Average ICU LOS for each Pleura's score and lung zone with respective confidence intervals obtained by bootstrap

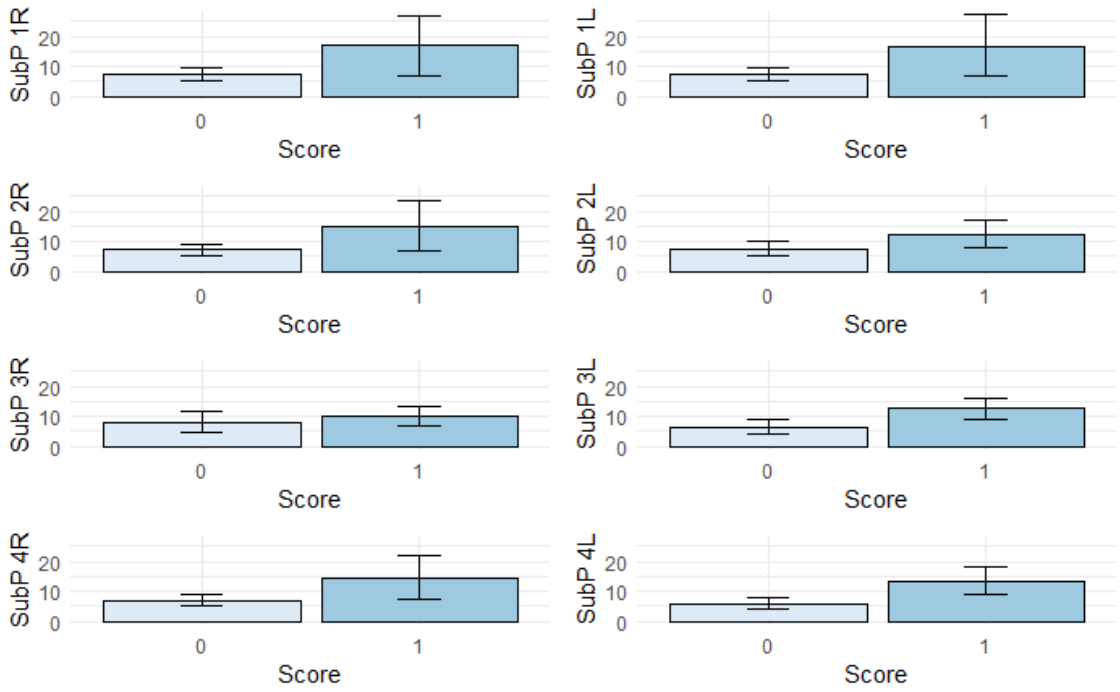

Figure 4.20: Average ICU LOS for each SubP's score and lung zone with respective confidence intervals obtained by bootstrap

the LOS on ICU. Finally, the Lobar Consolidations have, again, very unstable results, and

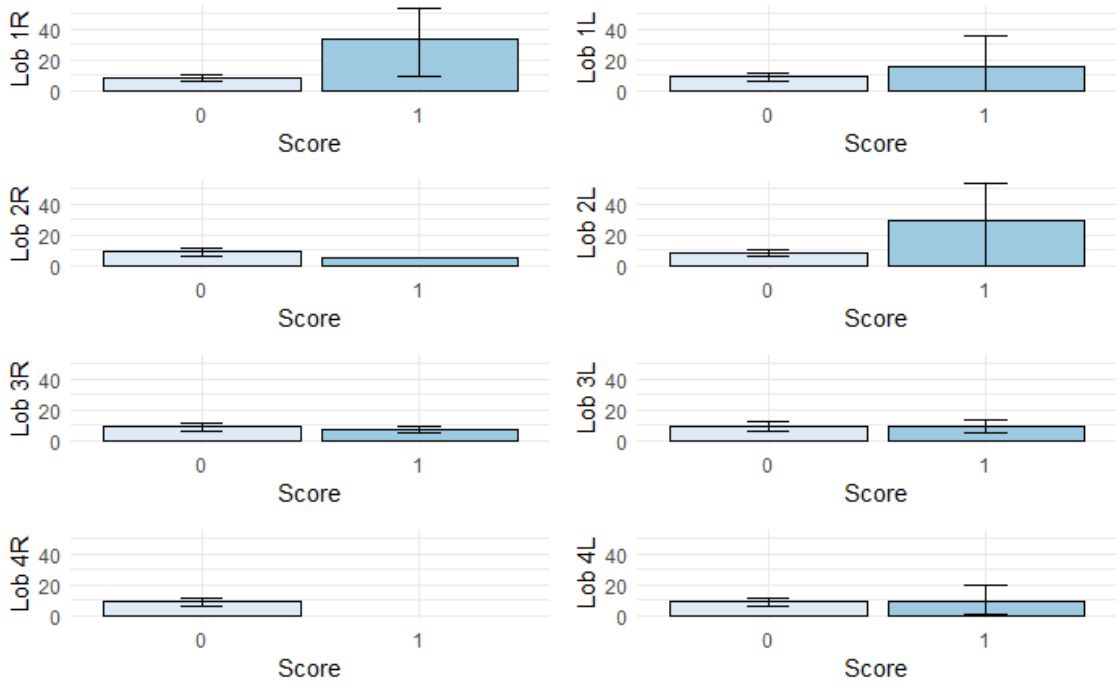

Figure 4.21: Average ICU LOS for each Lob's score and lung zone with respective confidence intervals obtained by bootstrap

few statistical conclusions can be taken, as shown in figure 4.21. The main conclusion is that these consolidations tend to aggravate the patient status.

4.1.4 Conclusions

In general, B-lines is the variable whose scores can better explain mortality and LOS. There can be seen important evolution on Mortality, LOS and ICU LOS with the increase of the scores, and only the relation between score 2 and score 3 is more complex to explain given the influence mortality has on the other two variables (LOS and ICU LOS). SubP also has a straightforward interpretation as its presence increases all the mentioned variables. As the incidence is not excessively unbalanced, around 23% when considering the mean of all evaluated zones, there is enough information to make some inferences. On the other side, both Pleura and Lob have extremely unbalanced classes, which lead to worse and less robust conclusions, especially in the Lob case.

4.2 Severity indicator

The GLM's main goal is to construct a model where the independent variables can explain the response variable as well as possible. Therefore, the desired response variable is of primal importance. In logistic regression, a particular case of GLM's, the response variable is binary and is commonly used to model death or ICU hospitalizations as there are only two different outcomes. The LOS is also a variable that can be model through GLM's, but there are some details that must be taken under consideration. The most important one is that there are no negative values in LOS, which immediately makes linear regression a risky choice, as a negative coefficient may lead to negative predictions. The severity indicator must reflect all these variables, which often brings some challenges on deciding how to construct this response variable.
To construct the indicator three base ideas were applied:

- Step 1

  The basis of this indicator is the LOS, as, generally speaking, the longer someone stays in the hospital, the most severe is his case. This modelling decision has one big problem, as it assumes that someone that is hospitalized, for example, 10 days has a more severe disease than someone that dies in 5 days, which is ultimately a big flaw, as it is not realistic.

- Step 2

  To solve the issue pointed in step 1, a big M approach was taken. This means that every patient that passed away instead of having its LOS indicating the severity,

it has a value "M", bigger than the LOS of all patient's that lived. This approach leads to the severity of the deceased being always greater than of those who lived. However it still may have some deficiencies. For example, if a patient has a mild case and leaves the hospital after 10 days it's severity indicator is equal to the Severity Index (SI) of someone who had a severe case and was hospitalized also 10 days but 8 of which in ICU.

- Step 3

  To make the impact of ICU hospitalization more clear on the indicator, the SI of the alive patients, instead of being only the LOS, is the LOS plus the number of days in ICU. This approach is applied only on the alive patients, as in some cases the patients reach the hospital in such a severe status that ICU hospitalization is not suited for them.

Until now, all assessments were made with patients in the center of the analysis. However, the purpose of this work is to create a link between the results of lung ultrasounds and the severity of the disease of a given patient. Consequently, it is pertinent to note that the status of the patients evolve during the hospitalization, which will be reflected on the lung ultrasounds. Therefore, the ultrasounds associated to a patient must also reflect this evolution which requires that the SI should be adapted to reflect it.
The following adaptations explain how the initial steps were altered to better suit the purpose of the indicator:

- Adaptation 1

  The first adaptation from the patient's SI is that instead of using the LOS, it is used the time until discharge, given the date of each ultrasound, in the case of alive patients. This approach is based on the idea that the closer a patient is to discharge day, the less severe is its status.

- Adaptation 2

  On a similar line of thought, in the case of patients that died, the closer they are to the day that they passed, the more severe is their status. Therefore, along with the big M approach, from this M one subtracts the number of days until death, in accordance to the day of each ultrasound. This way, the closer to the death, the bigger is the SI.

- Adaptation 3

  In the same way of thinking, instead of adding the number of ICU days to a patient's SI, it is added the number of days until ICU discharge.

These adaptations led to the construction of 6 different scores, each integrating a part of the ideas presented:

- M - Uses only the big M approach, M for dead patients and Time Until Discharge (TUD) for the others.

$$\begin{cases} \text{TUD if Alive} \\ \text{M if Dead} \end{cases}. \tag{4.1}$$

- M_H - Uses the big M subtracted by the Time until Death (TD) for dead patients and TUD for the others.

$$\begin{cases} \text{TUD if Alive} \\ \text{M-TD if Dead} \end{cases}. \tag{4.2}$$

- M_I - Uses the big M for dead patients and TUD added with Time Until ICU Discharge (TUID) for the others.

$$\begin{cases} \text{TUD+TUID if Alive} \\ \text{M if Dead} \end{cases}. \tag{4.3}$$

- M_H_I - Combines M_H and M_I

$$\begin{cases} \text{TUD+TUID if Alive} \\ \text{M-TD if Dead} \end{cases}. \tag{4.4}$$

- M_H_aI - Adds a coefficient ($a \in [0, 0.9]$) to the TUID

$$\begin{cases} \text{TUD+a}\times\text{TUID if Alive} \\ \text{M-TD if Dead} \end{cases}. \tag{4.5}$$

- M_bH_aI - Adds a coefficient ($b \in [0, 0.9]$) to the TD

$$\begin{cases} \text{TUD+a}\times\text{TUID if Alive} \\ \text{M-b}\times\text{TD if Dead} \end{cases}. \tag{4.6}$$

In all scores, the value of "M" must be chosen and, in the last two, "a" and "b" also have to be decided. The rule used for "M" is that it must be bigger than the maximum LOS of all patients, plus a small margin. The best values for coefficients "a" and "b" were studied varying them between 0 and 0.9 by intervals of 0.1. The values selected were those that maximized the KS test's p-value when the Scores were fitted to the Gamma or Log-normal distributions. Therefore, in total, there are 12 different scores, two for each

of the scores described, one optimized to best approximate the Gamma distribution and other optimized with respect to the Log-normal Distribution.

4.2.1 Scores Gamma

In figure 4.22 are the plots of the scores described, with the Gamma distribution theoretical curve for comparison, and the respective p-value for the KS test. Table 4.1 summarizes the values that optimize the approximation. The table's values show, as it was expected,

Table 4.1: Optimal values for M, a and b, in the different Gamma scores, in regard to the KS test's p-value

Scores	M	a	b
M	76	-	-
M_H	84	-	-
M_I	76	-	-
M_H_I	76	-	-
M_H_aI	76	0.9	-
M_bH_aI	76	0.3	0.9

that the smaller the "M" value the better is the fit to the Gamma distribution. In fact, the maximum LOS was 72 days, and to give a little margin, 76 was indeed the smaller value allowed for "M". In respect to the "a" coefficient, in the first case it's 0.9, the maximum value allowed, which leads to the conclusion that a larger value of "a" would lead to a better fit. This is corroborated by the score M_H_I, that ultimately is M_H_aI with a=1, which has a higher p-value. Finally, in the last score the value of "b" is also the largest possible, although the value of "a" is smaller. The differences between these last two scores and M_H_I are small and do not lead to higher p-values upgrades, therefore being worse options.

In the plots shown in figure 4.22, the effects of adding M, TD and TUID can be better visualized. Obviously, in the simplest score M, the big M approach in dead patients leads to a high bar that clearly degrades the quality of the fit, although the p-value is still acceptable. In M_H, it can be seen that, by subtracting TD from M, the curve flattens and the p-value rises substantially. The effect of adding the TUID to TUD in alive patients is equivalent to say that ICU days count as double of a "normal" nursery hospitalization day. The effect of this idea is that it leads to less small values and the distribution, while still right skewed, has smaller differences between the first bars and the middle ones. The p-value, however, decreases from the original (first) score. The score M_H_I is created by joining the two previous ideas, and the explained effects seem to blend well together, leading to the highest p-value of all scores. As mentioned earlier, the adding of coefficients doesn't create much value in relation to the goodness of fit.
The main goal, however, is not to have an indicator that has an approximate Gamma distribution but one that is a good tool to separate the patients with severe complications

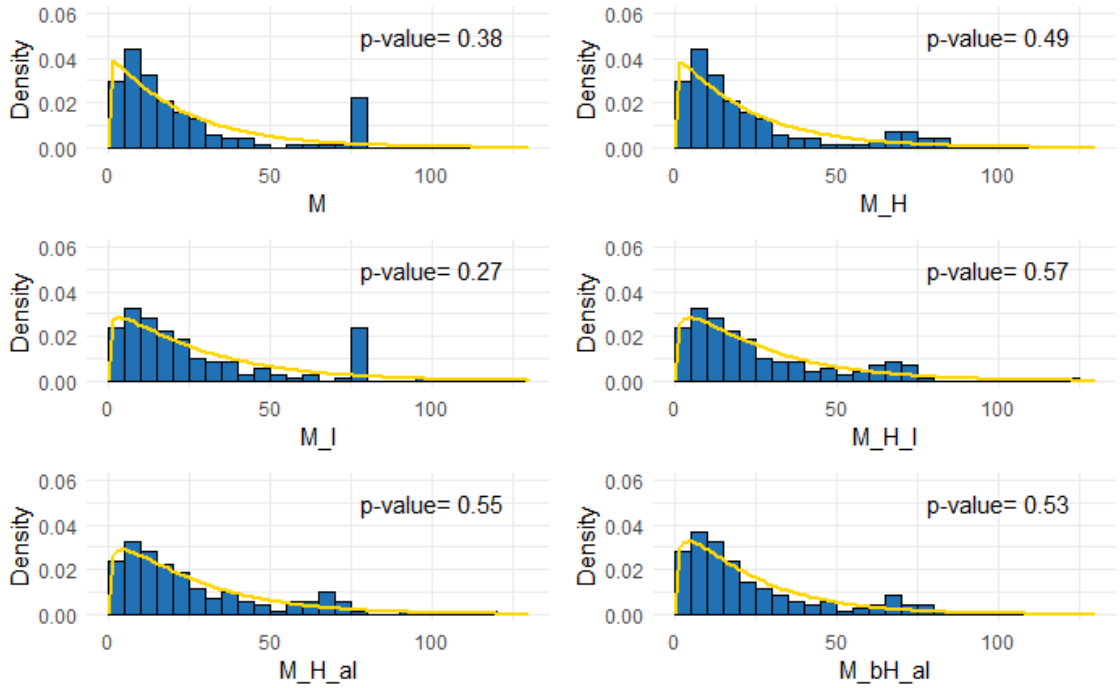

Figure 4.22: Gamma scores distribution with estimated Gamma parent distribution and respective KS test's p-value

from those with less problems. To assess the quality of these scores, Receiver Operating Characteristic (ROC) curves are used. Figure 4.23 shows the results in regard to the capacity of separating the patients that died from those that didn't. The scores' ability of separating the patients that needed ICU hospitalization from those that didn't, are in figure 4.24 (only the patients that didn't die were taken in consideration in this case). And finally, figure 4.25 shows the scores' capacity of differentiating patients that died or had ICU hospitalizations from those that stayed in the nursery the whole time and didn't die.

As expected, the big M approach leads to a high Area Under the Curve (AUC) for all scores when predicting death. This is good but doesn't allow to differentiate the scores between them, and therefore doesn't help in choosing the best score to use.

The ICU hospitalization prediction shows significant differences between scores, but also points some similarities between them. To start, the prediction doesn't have as good results as in the death case but nonetheless, M_I and M_H_I have a good AUC even noting the large confidence intervals. The next best score is M_H_aI which ultimately is M_H_I with a slightly less influential TUID, and that explains the slightly smaller capacity to differentiate ICU hospitalizations. M_H_aI is followed by M_bH_aI, again, as "a" is smaller than 1 and 0.9, it´s natural that the predicting ability is worse when compared to M_H_I and M_H_aI. Lastly, the first two scores are the worst at predicting ICU hospitalization but still with a decent prediction. The formed pairs can be explained by the small sample

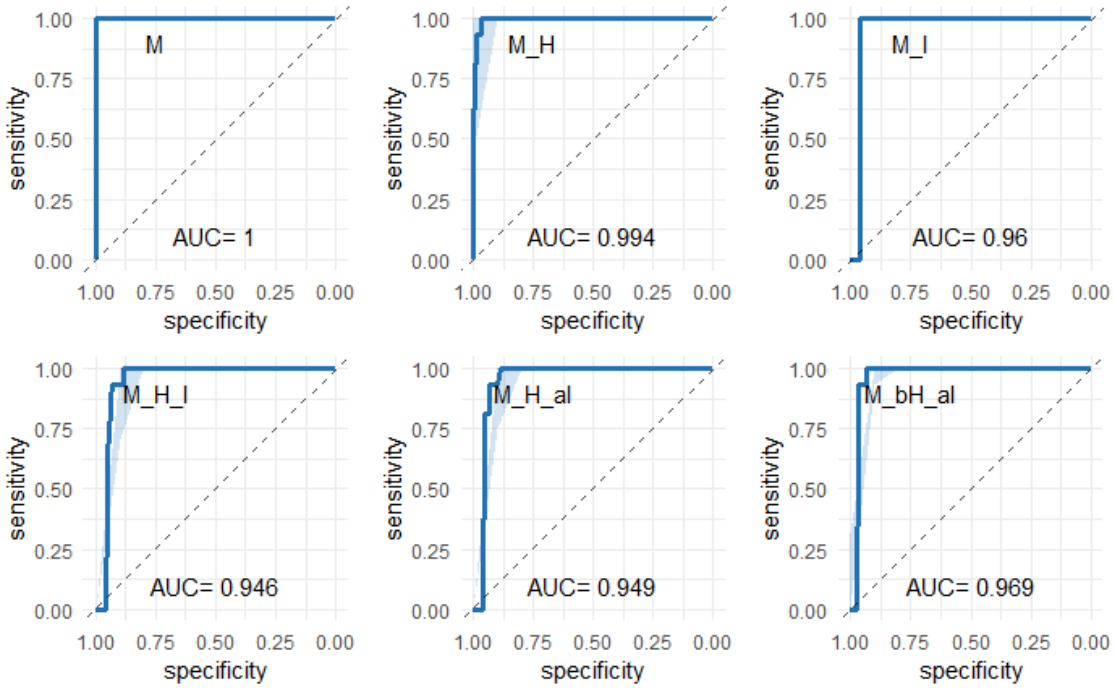

Figure 4.23: Gamma scores ROC curves for predicting death with respective confidence interval obtained by bootstrap

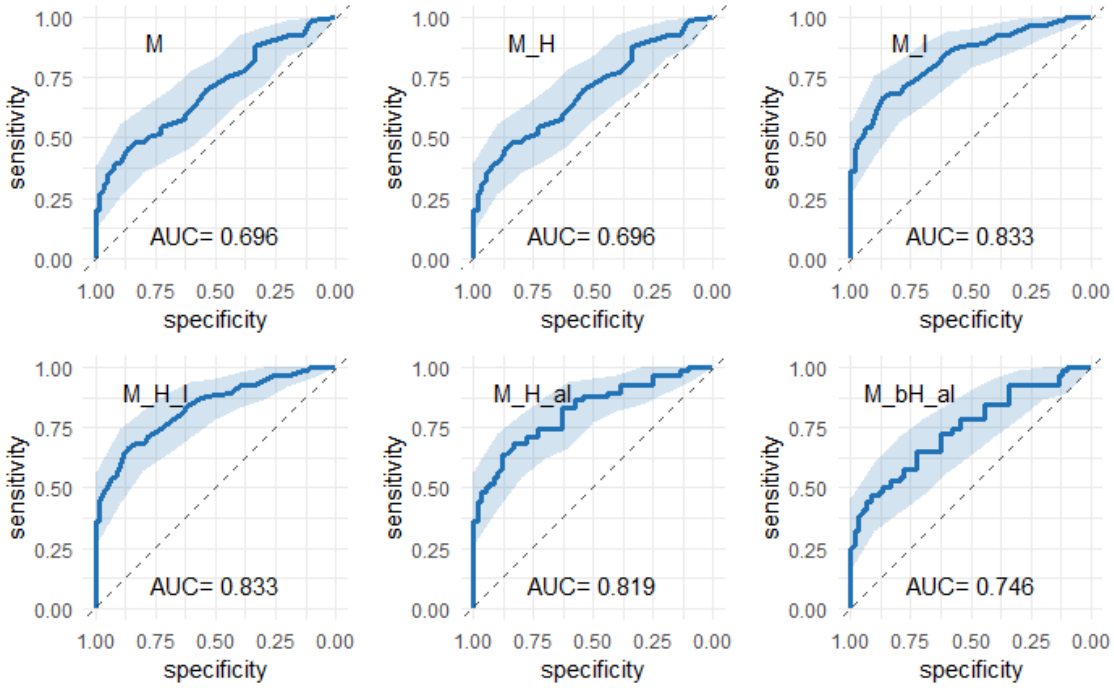

Figure 4.24: Gamma scores ROC curves for predicting ICU hospitalization with respective confidence interval obtained by bootstrap

of patients that died. This leads to the TD variable being applied in a very small numbers of lung ultrasounds and, therefore, conducts to almost no differences between M and M_H and also from M_I and M_H_I.

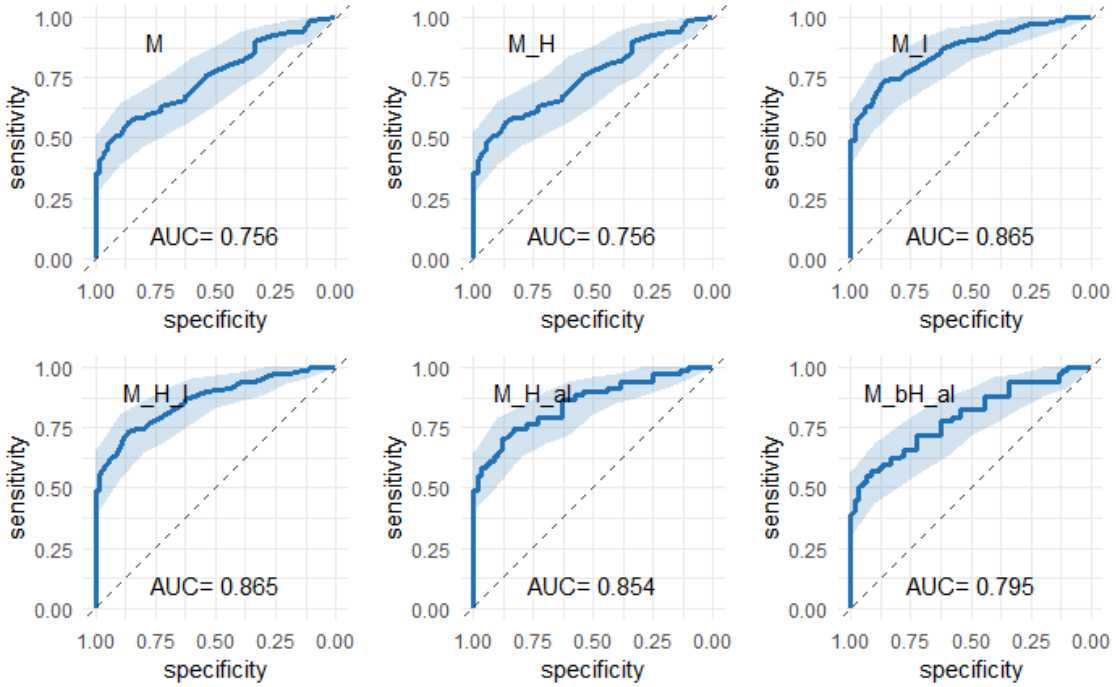

Figure 4.25: Gamma scores ROC curves for predicting ICU hospitalization or death with respective confidence interval obtained by bootstrap

Finally, the joint prediction show good results in all scores, helped by the almost optimal prediction of death, and the conclusions taken are similar to the ICU case.
Taking all this information into account, the scores with best predicting ability are M_H_I and M_I, but given that M_H_I has a better fit to the Gamma distribution, this is the one chosen to be the Gamma Severity Index.

4.2.2 Scores Log-normal

A similar process was executed for score optimization regarding the Log-normal distribution. In table 4.2 are presented the respective best values of "M", "a" and "b" in the scores. In figure 4.26, are the correspondent plots and the Log-normal theoretical distributions.

Table 4.2 shows a different behaviour when compared to the Gamma scores although 76 is also the minimum value accepted for M. The values of "M" are significantly higher and in two cases even reach the higher value possible, 120. On the other hand, the coefficients "a" and "b" have smaller values than in the Gamma case, making the last two scores markedly different from M_H_I (table 4.2).

Table 4.2: Best values for M, a and b, in the different Log-normal scores, in regard to the KS test's p-value

Scores	M	a	b
M	97	-	-
M_H	120	-	-
M_I	92	-	-
M_H_I	120	-	-
M_H_aI	106	0.1	-
M_bH_aI	96	0.1	0.2

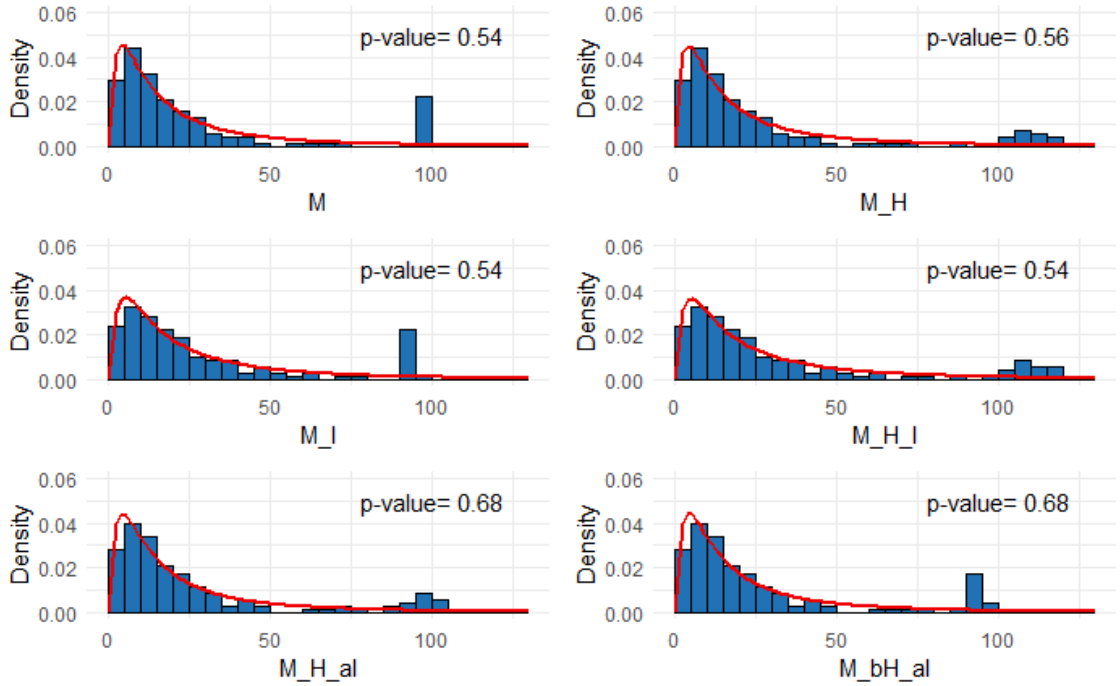

Figure 4.26: Log-normal scores distribution with estimated Log-normal parent distribution and respective KS test's p-value

In figure 4.26 are shown the plots regarding the Log-normal scores with the respective parent distribution curve. The main differences, when comparing them with the Gamma scores, are in the largest values. As the value of "M" is larger there is a bigger difference between the dead patients' scores and the alive ones. In the first four scores the only difference between the Log-normal scores and the Gamma ones is exactly the value of "M". As the coefficients values are small their use leads to a small effect of ICU hospitalization on the severity.

Assessing now the quality of theses scores prediction wise, the following figures show the ability to predict death (figure 4.27), ICU hospitalization for patients that left the hospital alive (figure 4.28) and death or ICU vs alive and only nursery hospitalization (figure 4.29). The death prediction is, as it was on the Gamma case, almost perfect, because the big M

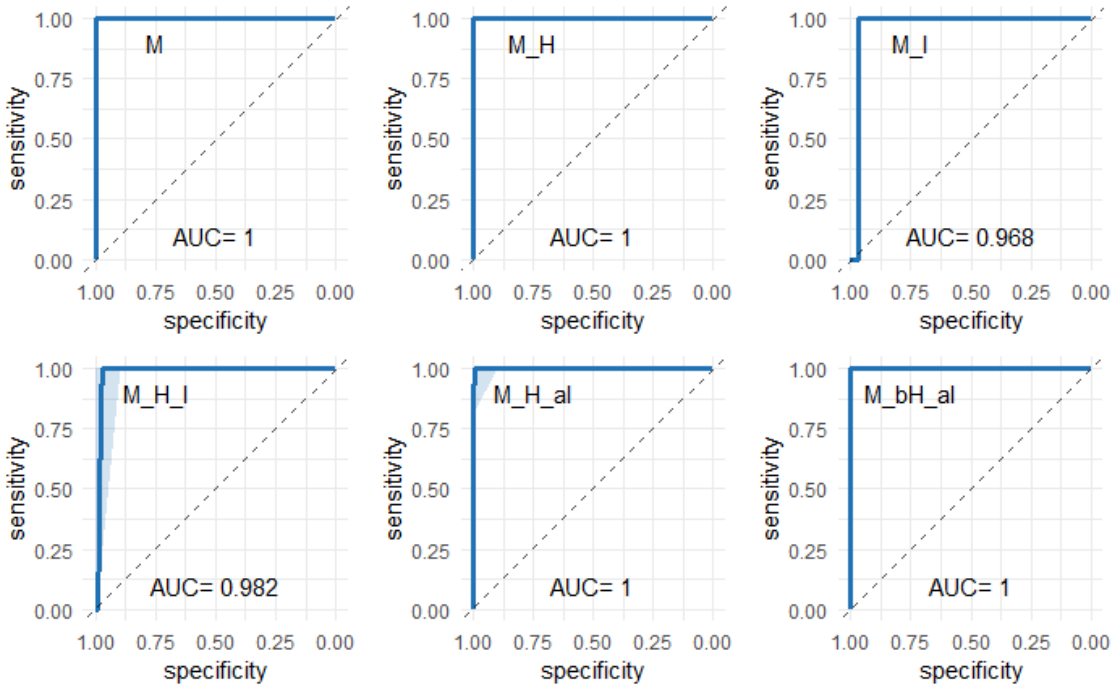

Figure 4.27: Log-normal scores ROC curves for predicting death with respective confidence interval obtained by bootstrap

effect separates almost flawlessly the dead patients from the alive ones (figure 4.27).

The ability to separate the alive patients that needed ICU hospitalization, from those that didn't, is equal to the Gamma case in the first four scores, because the only difference between them is the value of "M". As in the alive patients "M" doesn't have an influence, the values of AUC are exactly the same. In the Log-normal scores with coefficients there is a smaller influence o TUID, the separating ability of them is considerably smaller when compared to the Gamma case. Finally, when predicting ICU or Death, the first four scores have the exact same AUC of the Gamma cases, which leads to the same conclusion, these scores are reasonably good at making these separations. The only difference in the scores with coefficients, as already referred, is that they have a smaller capability of predicting ICU hospitalization, and therefore present worse results.

To conclude, having in consideration the goodness of fit and the quality of prediction, M_I and M_H_I are the best scores. As M_H_I has more information and differentiate the dead patients among themselves, it's the chosen for use in the regression.

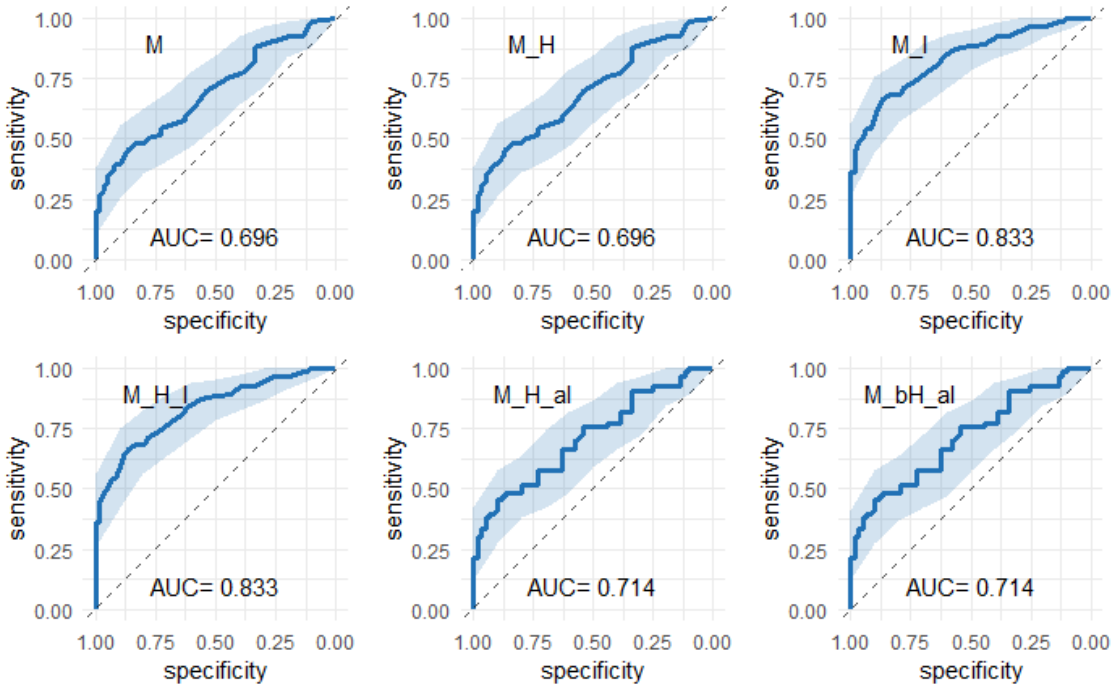

Figure 4.28: Log-normal scores ROC curves for predicting ICU hospitalization with respective confidence interval obtained by bootstrap

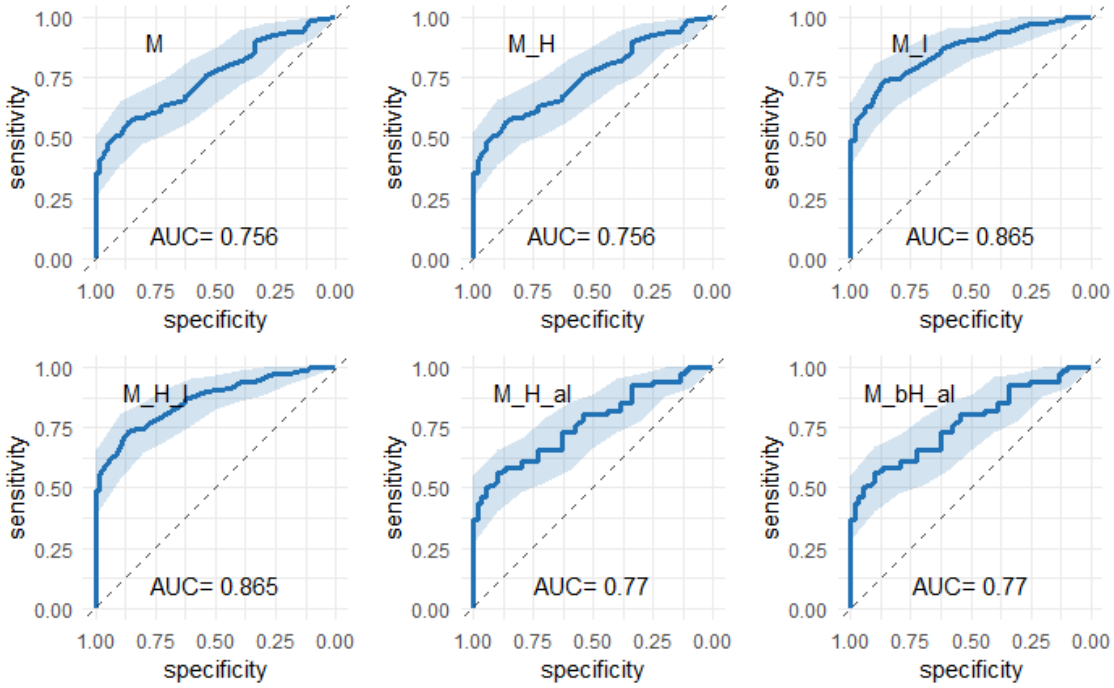

Figure 4.29: Log-normal scores ROC curves for predicting ICU hospitalization or death with respective confidence interval obtained by bootstrap

4.3 Principal Components Analysis

Principal Components Analysis (PCA) is typically used to reduce the number of variables when variables have high correlations. Principal Components Analysis by means of Alternating Least Squares (PRINCALS) is a technique with the same purpose, but more suited for ordinal variables.

4.3.1 Correlations

As there appeared to be some correlations between the same variables in different lung zones, a correlation analysis was made with using Kendall's τ because the variables are ordinal. In table 4.3 one can see the Kendall's correlations of the B-lines variables in the

Table 4.3: Kendall's correlations of B-lines' scores in each lung zone

	1R	2R	3R	4R	1L	2L	3L	4L
1R	1	0.58	0.4	0.48	0.45	0.51	0.45	0.49
2R	0.58	1	0.51	0.45	0.47	0.51	0.49	0.41
3R	0.4	0.51	1	0.7	0.54	0.54	0.67	0.57
4R	0.48	0.45	0.7	1	0.52	0.49	0.52	0.5
1L	0.45	0.47	0.54	0.52	1	0.66	0.54	0.47
2L	0.51	0.51	0.54	0.49	0.66	1	0.61	0.59
3L	0.45	0.49	0.67	0.52	0.54	0.61	1	0.69
4L	0.49	0.41	0.57	0.5	0.47	0.59	0.69	1

different lung zones. All the correlations are positive, which indicates that the severity in a given zone is, in general, in concordance with the other zones. The majority of the values are above 0.5, and the minimum correlation value is 0.4. Although the correlations aren't excessively high, there are clearly patterns that can be simplified.

In the case of Pleura, the correlations are less evident (table 4.4). There are only a minority of pairs where the correlation is superior to 0.5 and the majority is situated between 0.25 and 0.5.

In table 4.5 are the correlations related to the SubP variable, which is where correlations really start to fade. The largest value is 0.44, but the majority of them is under 0.25. This indicates that PCA will not be as effective as there will be the need of having a large

Table 4.4: Kendall's correlations of Pleura's scores in each lung zone

	1R	2R	3R	4R	1L	2L	3L	4L
1R	1	0.45	0.14	0.28	0.49	0.27	0.09	0.41
2R	0.45	1	0.47	0.26	0.25	0.42	0.34	0.33
3R	0.14	0.47	1	0.27	0.18	0.33	0.61	0.25
4R	0.28	0.26	0.27	1	0.37	0.31	0.19	0.64
1L	0.49	0.25	0.18	0.37	1	0.53	0.06	0.41
2L	0.27	0.42	0.33	0.31	0.53	1	0.3	0.36
3L	0.09	0.34	0.61	0.19	0.06	0.3	1	0.42
4L	0.41	0.33	0.25	0.64	0.41	0.36	0.42	1

Table 4.5: Kendall's correlations on SubP's scores in each lung zone

	1R	2R	3R	4R	1L	2L	3L	4L
1R	1	0.38	0.11	0.33	0.06	0.17	0.04	0.02
2R	0.38	1	0.28	0.3	0.13	0.13	0.17	0.15
3R	0.11	0.28	1	0.35	0.11	0.44	0.36	0.23
4R	0.33	0.3	0.35	1	0.19	0.34	0.17	0.32
1L	0.06	0.13	0.11	0.19	1	0.35	0.13	0.15
2L	0.17	0.13	0.44	0.34	0.35	1	0.11	0.37
3L	0.04	0.17	0.36	0.17	0.13	0.11	1	0.4
4L	0.02	0.15	0.23	0.32	0.15	0.37	0.4	1

number of components to explain the variance in the data.

Finally, in the Lob case, table 4.6, the small sample of lung ultrasounds with this indicator present leads to very erratic correlations. First of all, as in zone 4R there are no positive

Table 4.6: Kendall's correlations on Lob's scores in each lung zone

	1R	2R	3R	4R	1L	2L	3L	4L
1R	1	-0.01	-0.02		-0.02	0.5	0.12	-0.03
2R	-0.01	1	0.57		-0.01	-0.02	0.25	-0.02
3R	-0.02	0.57	1		-0.02	-0.03	0.44	0.46
4R								
1L	-0.02	-0.01	-0.02		1	0.77	0.28	0.7
2L	0.5	-0.02	-0.03		0.77	1	0.19	0.53
3L	0.12	0.25	0.44		0.28	0.19	1	0.4
4L	-0.03	-0.02	0.46		0.7	0.53	0.4	1

cases of Lob, the value of correlation cannot be expressed. There are also cases where the correlation is negative, which would mean that a certain zone being unhealthy leads to other zones being healthy. This just shows that this correlation values can't be taken in consideration on the decision of using PCA or not.
The final decision was to use PCA as the variable with most variability, the B-lines, may benefit from this approach. Although the binary variables have less correlation between lung zones, the use of more Principal Component (PC) will certainly rectify this.

4.3.2 Scree plots

After applying the PCA for ordinal data, PRINCALS, it's necessary to decide how many PC's will be chosen to be part of the new dataset. Usually, when PCA techniques are used, there are two indicators to decide how many PC's should be used. The first is the Scree plot, a plot with the PC's decreasingly ordered by the percentage of variance explained, with a line uniting the dots. The number of PC's chosen is the number where the line starts to be horizontal. The other method is to choose the number of PC's that explain between 70% and 90% of the total variance.
Figure 4.30 depicts the scree plot associated with the B-lines variables. It clearly shows, that the first PC is key to explain the variance of the data. Through the criteria explained before, two PC should be kept, as is where the line becomes horizontal, and at the same time, they explain 74% of the total variance.
Regarding the Pleura's variables, the variance is more spread along the PC's, as shown in figure 4.31. Again, the line becomes approximately horizontal on the second PC and 59%

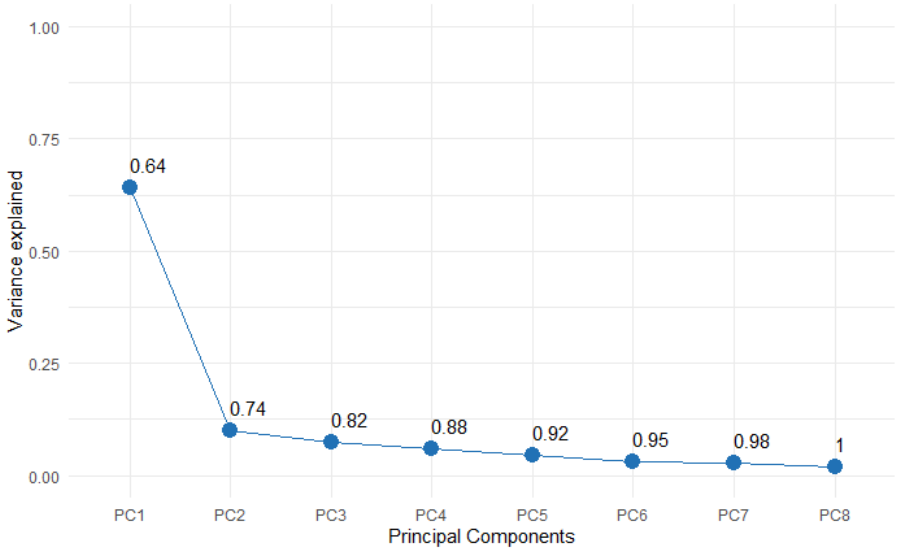

Figure 4.30: Scree plot of PRINCALS applied to B-lines' scores in each lung zone and the cumulative variance explained, in proportion, in each principal component

of the variance is an acceptable value. Therefore, two PC's were chosen regarding Pleura.

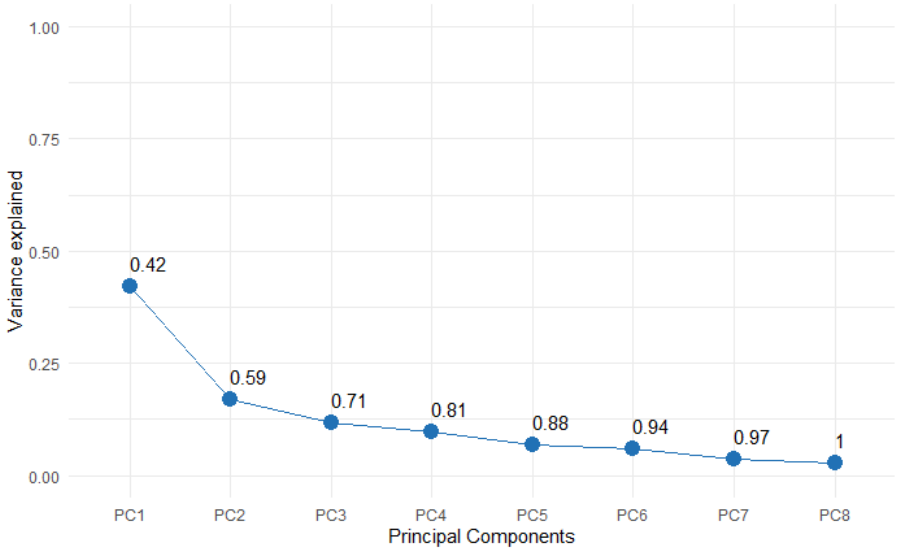

Figure 4.31: Scree plot of PRINCALS applied to Pleura's scores in each lung zone and the cumulative variance explained, in proportion, in each principal component

As was expected, when advancing through the variables, the correlations start to be less evident. This leads to the variance being more diluted through the PC's instead of being concentrated on the first two. This, in the case of SubP, leads to the first two PC's only explaining about 50%, figure 4.32. As this value is considerably far from the 70%, even tough the horizontal line starts on the second component, three PC's are kept.
The final variable, Lob, as mentioned previously, has the problem of not having any positive case on zone 4R. Therefore, only 7 PC's can be drawn (figure 4.33). This scree

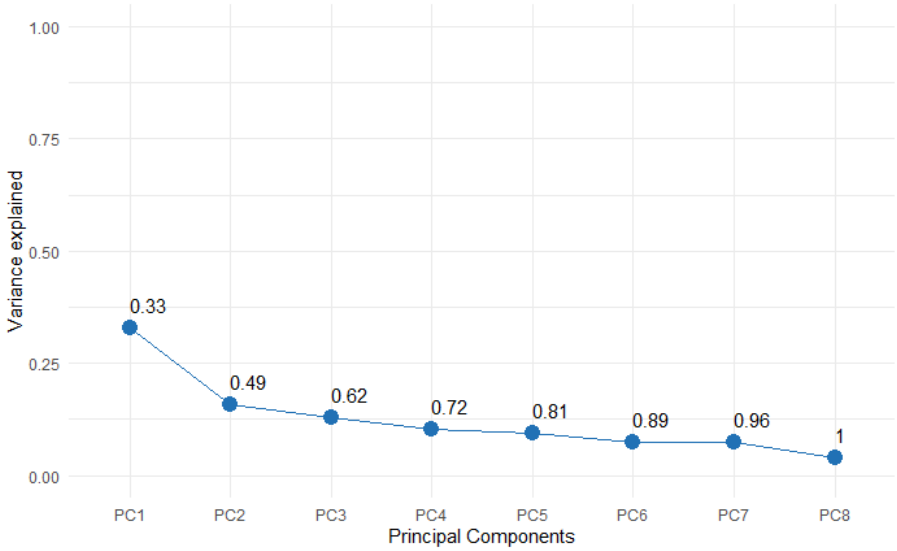

Figure 4.32: Scree plot of PRINCALS applied to SubP's scores in each lung zone and the cumulative variance explained, in proportion, in each principal component

plot is the only where the horizontal line starts only on the fourth PC. However, as 80% of the variance is explained by the first three, three is the number chosen.

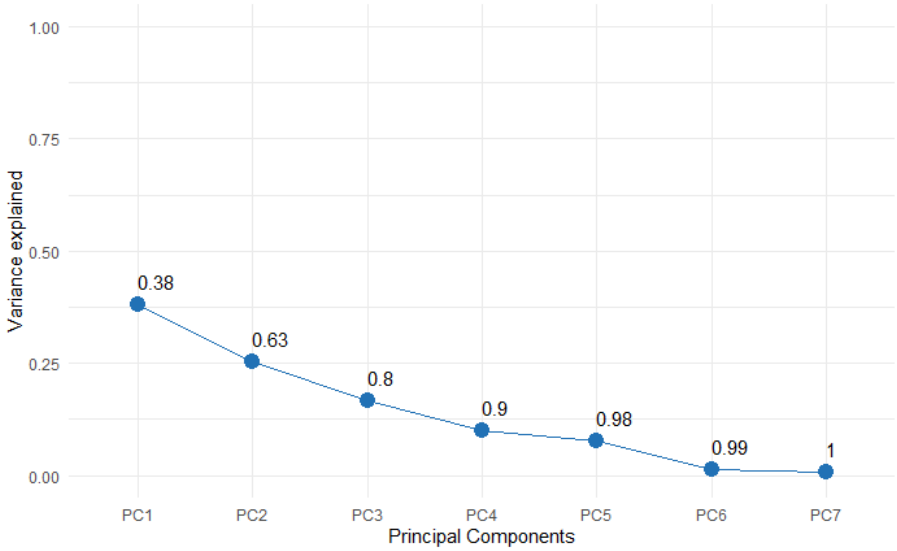

Figure 4.33: Scree plot of PRINCALS applied to Lob's scores in each lung zone and the cumulative variance explained, in proportion, in each principal component

Finally, on table 4.7, are the loadings associated with the respective lung zones and PC's, where "B1"represents the first PC of B-lines and "B2"represents the second PC. Similarly for the remaining variables. One important note is that, in all cases, the first PC has all loadings with the same signal and, with the exception of Lob, the values are all very close to each other. This can be interpreted as the first PC being used to assess the overall severity of the ultrasound. Besides this, there are no signs of the second, or other, PC separating the right zones from the left ones.

Table 4.7: Loadings of Principal Components of B-lines' scores (B1 and B2), Pleura's scores (P1 and P2), SubP's scores (S1, S2 and S3) and Lob's scores (L1, L2 and L3)

	B1	B2	P1	P2	S1	S2	S3	L1	L2	L3
1R	0.77	-0.49	0.60	0.42	0.41	-0.73	-0.03	0.24	-0.26	-0.90
2R	0.75	-0.49	0.68	-0.18	0.53	-0.52	-0.24	0.22	0.74	-0.25
3R	0.83	0.28	0.61	-0.61	0.68	0.13	-0.16	0.44	0.80	-0.03
4R	0.79	-0.08	0.65	0.22	0.68	-0.23	0.04	0.00	0.00	0.00
1L	0.80	-0.04	0.64	0.51	0.42	0.14	0.64	0.81	-0.38	0.30
2L	0.84	0.04	0.69	0.06	0.67	0.17	0.47	0.78	-0.49	-0.26
3L	0.84	0.34	0.57	-0.67	0.51	0.40	-0.54	0.60	0.40	-0.07
4L	0.79	0.35	0.75	0.15	0.60	0.43	-0.15	0.85	-0.01	0.35

4.4 GLM Regression

With the scores converted to its principal components and with the severity indicator chosen, it's now possible to make the regression. In addition to the lung ultrasounds information, age and gender are also used as variables.

The two versions of the severity indicator follow different distributions. Therefore, the regression model must be adapted to each case. In the Gamma case, there are three possible link functions identity, log or inverse. The identity link function is dangerous because it would turn into a regular linear regression, with the possibility of negative values. The inverse function presents the same problem. Only log guaranties that the predictive values are strictly positive. Regarding the score with Log-normal distribution, the idea is to transform the scores with the log function and then, as the scores will follow a normal distribution, apply GLM with the Gaussian family, which leads to the link function being the identity.

It is also important to note that the individual number 34 was withdrawn from the regression as it was considered an outlier. This individual was hospitalized for over two months and therefore had a very high score when compared to all the other patients. As the sample is relatively small the effect of outliers is increased.

In the Gamma model, as was previously mentioned, there are three possible link functions. Only two were used because the identity link will be used on the Log-normal scores. A comparison will be made between three variants of the regression. The first variant is using a Generalized Linear Model assuming the Scores follow a Gamma distribution with inverse link function (G-Inv); the second is a Generalized Linear Model assuming the Scores follow a Gamma distribution with log link function (G-Log); the final variant is a Generalized Linear Model assuming the logarithm of the Scores follow Normal distribution with identity link function (LN).

4.4.1 Gamma scores - Inverse Link function

The first step was to built the null model and the model with all the variables, so that the stepwise algorithms could be applied. The three algorithms (forward, backwards and

bidirectional) were applied to reach three final models, that, in this case were equal to each other.

Table 4.8 summarizes the model obtained by the stepwise algorithm. The results point

Table 4.8: Summary of the model obtained through the stepwise algorithm with selected variables' estimates, standard error and significance

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0.100	0.014	7.126	0.000
B1	-0.013	0.003	-4.346	0.000
Age	-0.001	0.000	-4.369	0.000
L1	-0.003	0.001	-2.578	0.011
Gender1	0.011	0.006	1.912	0.058
P2	0.005	0.003	1.543	0.125

that the more significant variables are Age and B1, as the p-value of the t-test is approximately 0, along with the intercept. After these two, L1 is the next variable significance wise, which still has a p-value that strongly encourages the use of it. The last two, Gender and P2, are the ones that lead to more doubts. Although the algorithms included them, the p-values indicate that their withdrawal may not affect the quality of the model, as the p-values are above 0.05, which means that the possibility of the respective coefficients being equal to zero is not rejected.

Before testing the effect of removing these two variables it's important to note that age has a negative estimate, which indicates that the increase of age leads to an increase of severity. This can be concluded from the fact that the linear predictor is on the denominator, so a decrease in it, leads to an increase on the overall predictor. In the same direction are B1 and L1, which is logical, as the first PC is an assessment of overall severity, the increase of this variable leads to a a decrease in the linear predictor and, consequently, to an increase in the overall prediction.

P2 is a variable with difficult interpretation, as the loadings of this PC don't appear to have a clear logic. This PC has negative loadings on zone 2R, 3R and 3L, along with a value very close to zero in zone 2L. A possible interpretation of this PC would be to differentiate zones 2 and 3 from zones 1 and 4. Assuming this, the positive estimate on the regression would mean that an increase of severity in zones 1 and 4 when compared to zones 2 and 3, meaning an increase in the PC related value, would mean an increase on the linear predictor and, therefore, a decrease on the overall severity. Summing up, an increase of severity on zones 2 and 3 when compared to zones 1 and 4 would lead to an increase on the overall severity. Finally, Gender has a positive estimate, which means that females (gender = 1) have less severe cases when compared to men.

It is also interesting to note that the first PC of Pleura (P1) is not included in the model, but P2 is. It would be expected that the PC that explain more variance would come first on variable selection.

To compare the performance of the model when removing the referred variables nested

models comparisons were made, using the F test. The first models compared was the one obtained after the stepwise algorithms (B1+Age+L1+Gender+P2) versus the same model without P2, as P2 is the variable with less significance. The F test resulted in a p-value of 0.14, which means that the model with more variables isn't clearly superior to the one without P2. Therefore, to acknowledge parsimony, the smaller model will be selected. Table 4.9 shows the new model.

The estimates didn't change, with the given rounding, but the p-values have slight vari-

Table 4.9: Summary of the model after removal of P2 with selected variables' estimates, standard error and significance

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0.100	0.014	6.908	0.000
B1	-0.013	0.003	-4.339	0.000
Age	-0.001	0.000	-4.247	0.000
L1	-0.003	0.001	-2.389	0.018
Gender1	0.011	0.006	1.820	0.071

ations. Again, Gender is not significant enough considering a significance of 5%, and therefore, another nested models comparison was performed, between the bigger model (B1+Age+L1+Gender) and the smaller model (B1+Age+L1). The test originated a p-value of 0.08, and maintaining 5% as the significance, the hypothesis that the estimate of Gender is 0 is not rejected. Therefore, the selected model is (B1+Age+L1). The final model

Table 4.10: Summary of the final model with selected variables' estimates, standard error and significance

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0.094	0.013	6.961	0.000
B1	-0.012	0.003	-4.153	0.000
Age	-0.001	0.000	-4.043	0.000
L1	-0.004	0.001	-3.564	0.001

has slight changes on the estimates and the significances this time don't leave any doubt, as can be observed in table 4.10.

It's also important to make the residuals analysis, as it may uncover some deficiencies of the model. The first plot, figure 4.34, has the deviance residuals printed vs a transformation of the fitted values, $2log(\hat{\mu})$, this transformation is advised by Turkman et al. [32]. The residuals are centered around 0 and the variance is approximately the same when X values are between 5 and 7. After that, there are less points, which are closer to zero, leading to a variance decrease. However, it can be said that the plot doesn't have any sort of trend, and therefore the model appears to fit the data well.

Another important feature of residuals analysis is to see of the Pearson's residuals follow a normal distribution. This can be visualized by a QQ-plot of the values and the theoretical normal distribution quantiles, figure 4.35. The plot shows that although the tails have some dissident values, especially in the right tail, the center of the residuals is close to

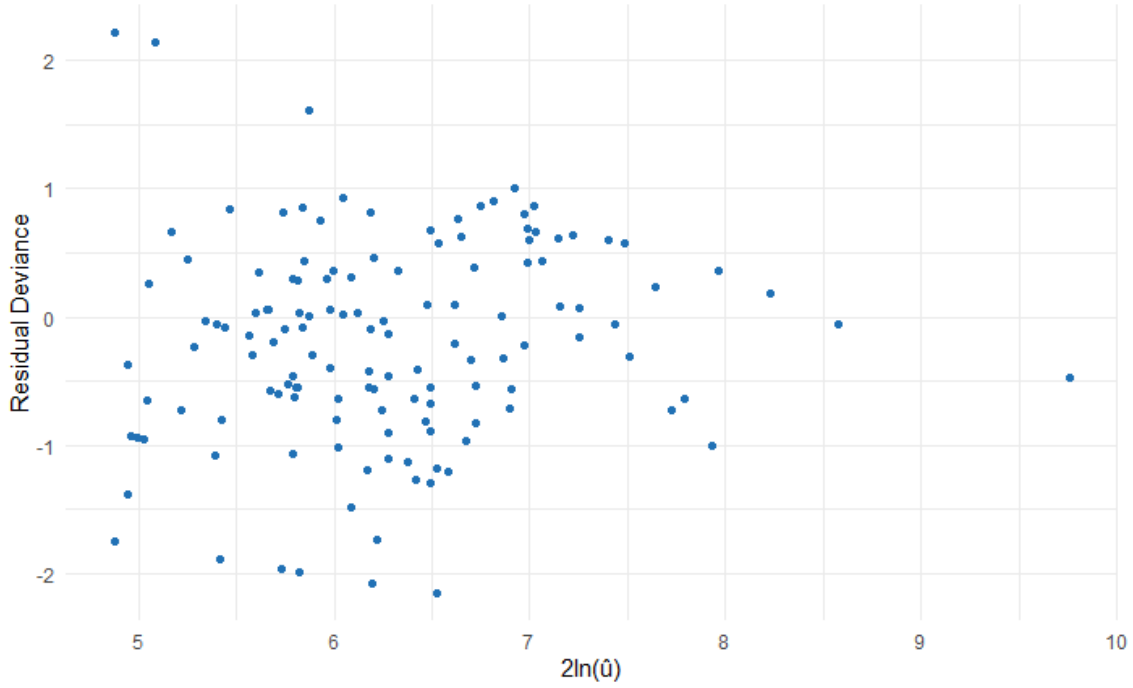

Figure 4.34: residual deviance of the final model vs $2log(\hat{\mu})$ where $\hat{\mu}$ are the model's fitted values

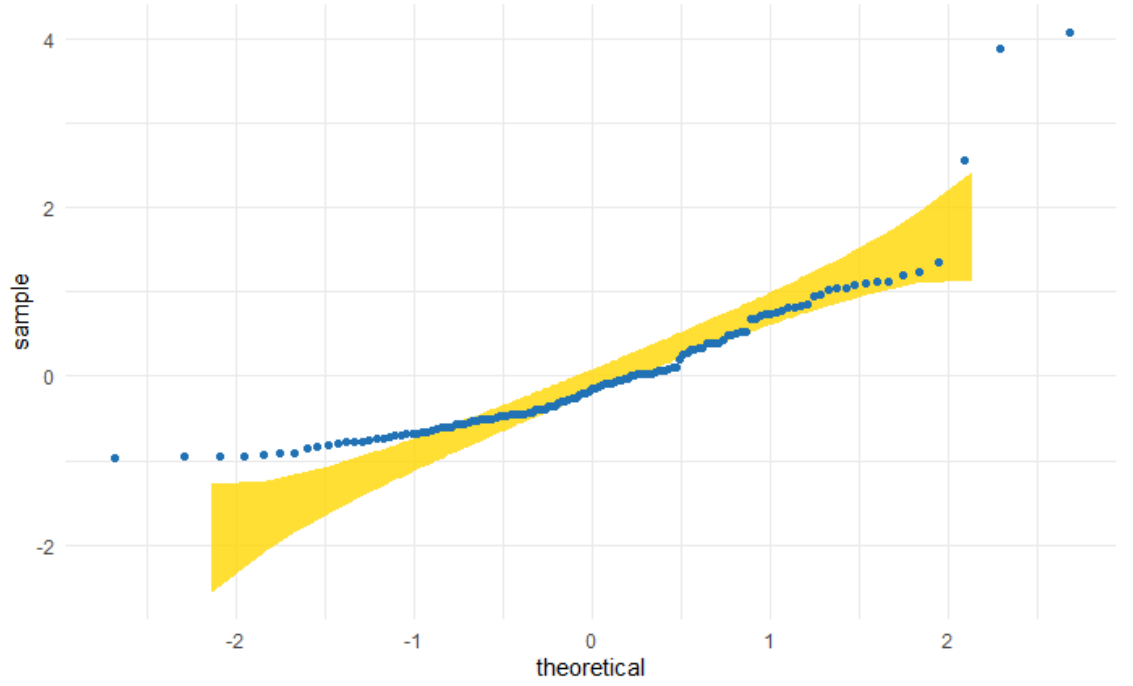

Figure 4.35: Pearson's residuals QQ plot compared to the normal distribution's quantiles

the theoretical line. Consequently, even though a goodness of fit test will probably reject normality, the Pearson's residuals can be considered as satisfactory.

Finally, to test the quality of the link function choice, a plot of the linear predictor vs the working response, $\hat{\eta} + \hat{D}(y - \hat{\mu})$, must show points as approximately a straight line. Figure

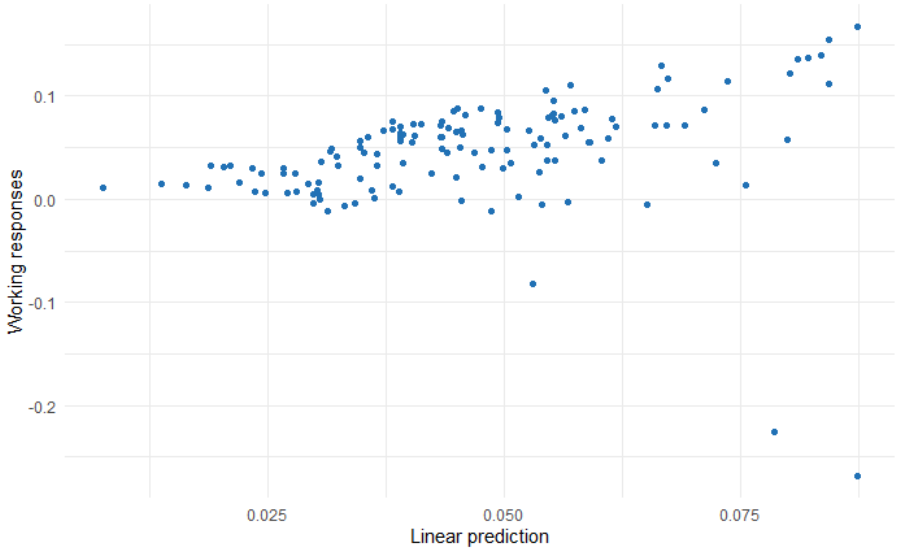

Figure 4.36: Final model's linear predictor vs working response

4.36, related to the plot referred, is the most flawed. The points can be considered to follow a relatively linear trend, but as the linear predictor increases the dispersion starts to grow. This may indicate that the inverse link may not be ideal to fit the data.

4.4.2 Gamma scores - Log link function

A similar approach was made using the Log as link function. All stepwise algorithms, again, reached the same model, however it is different than the one obtained using the inverse link. Table 4.11 has the information necessary to make a first analysis of the

Table 4.11: Summary of the model obtained through the stepwise algorithm with selected variables' estimates, standard error and significance

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	2.010	0.307	6.554	0.000
B1	0.244	0.084	2.890	0.005
Age	0.019	0.005	3.829	0.000
L1	0.162	0.072	2.258	0.026
S1	0.163	0.087	1.872	0.063
P1	-0.125	0.072	-1.736	0.085
P2	-0.097	0.069	-1.407	0.162
Gender1	-0.243	0.158	-1.539	0.126

model. The first three variables (and the Intercept) are the most significant (B1, Age and L1), with the respective p-values under 0.05 (which are the variables of the final model of the previous subsection). After these, S1 and P1 are the intermediate cases, because the

p-value is between 0.05 and 0.10, which leaves certain doubts about the importance of their introduction on the final model. The final two, P2 and Gender, have p-values above 0.10 which indicates that their removal may not damage the model's efficiency.
In regard of the estimates, B1, Age, L1 and S1 have positive coefficients and, therefore, the bigger the values the more severe are the cases. This is logical, as an older patient, in general, tends to have more severe consequences. In terms of B1, L1 and S1, as was previously mentioned, their estimates indicate the overall severity of all zones, so it was expected that the coefficients would be positive. P1, on the other hand has a negative estimate, what would mean that having heterogeneous pleura would be better than having normal pleura. This estimate, besides the existing doubts on the significance, is clearly distorted by the fact that only a small number of cases had homogeneous pleura. Again, second PC's are harder to interpret, given the nature of the loadings, but an interpretation similar to the previous subsection can be made. Gender's negative estimate indicates that women have less severe cases than men.
After the initial scrutiny, it's important to assess the performance of the model without the (possibly) non-significant variables. The first test was made to compare the stepwise model with the one obtained when P2, the less significant variable, is removed. The p-value of the F test was 0.13, and therefore the smaller model has a similar performance and should be chosen. Table 4.12 presents the new estimates and their corresponding

Table 4.12: Summary of the model after removal of P2 with selected variables' estimates, standard error and significance

	Estimate	Std. Error	t value	Pr($<\|t\|$)
(Intercept)	2.050	0.317	6.466	0.000
B1	0.229	0.087	2.632	0.010
Age	0.019	0.005	3.588	0.000
L1	0.162	0.074	2.184	0.031
S1	0.177	0.089	1.981	0.050
P1	-0.129	0.074	-1.735	0.085
Gender1	-0.231	0.163	-1.414	0.160

significances, which suffer small variations. The significance of the S1 decreases to 0.5, exactly the threshold of significance, and Gender becomes less significant. Another nested models comparison is performed, this time between this latter model and the one where Gender is removed. The p-value obtained is even higher in this case, 0.17, which again leads to the choice of removing Gender. The estimates of the new model suffered minor variations, as seen in table 4.13, and only P1 has a p-value above 0.05. This value leads to doubts of what the "correct" decision concerning the inclusion of P1 is, but, in the end, the chosen model was (B1+Age+L1+S1+P1) as having all the first PC's of each variable seemed logical.
The residuals analysis is similar to the previous subsection. The first plot, figure 4.37, shows the deviance residuals against a transformation of the fitted values. The values are

Table 4.13: Summary of the final model with selected variables' estimates, standard error and significance

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	2.137	0.309	6.915	0.000
B1	0.218	0.086	2.547	0.012
Age	0.016	0.005	3.329	0.001
L1	0.175	0.073	2.401	0.018
S1	0.177	0.088	2.012	0.046
P1	-0.139	0.074	-1.889	0.061

centered around zero. The variance appears to be constant as the x values increase, and, therefore, no objections are made from this plot.

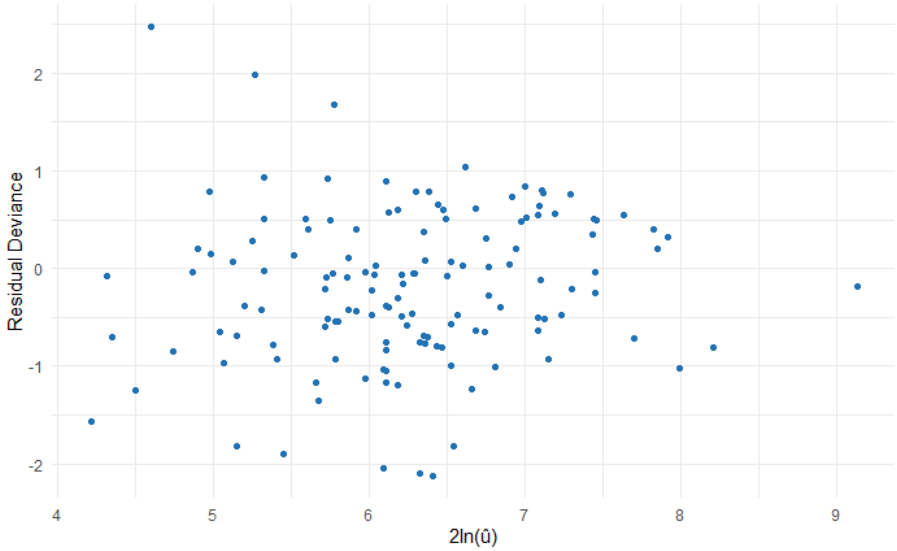

Figure 4.37: Final model's residual deviance vs $2log(\hat{\mu})$ where $\hat{\mu}$ are the model's fitted values

The next plot, figure 4.38, is the QQ plot of the Pearson's residuals, which are intended to have a normal distribution. As can be seen, the values in the middle are relatively close to the theoretical line and are within the confidence interval. In the tails, however, a few points are out of this interval. As previously mentioned, Pearson's residuals sometimes have problems with right skewed distributions, such as Gamma, and these results don't mean necessarily that the model isn't well fitted.

Finally, in figure 4.39, there is the plot which is intended to test the quality of the link function. If the choice is right, then the points should be distributed around a straight line. In this case, that line is not clear, although it appears to have a linear trend.

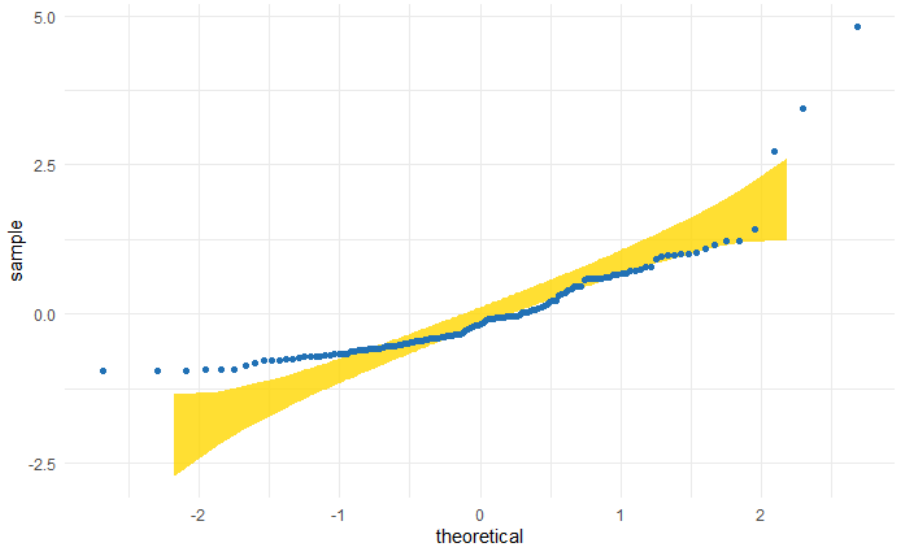

Figure 4.38: Pearson's residuals QQ plot compared to the normal distribution's quantiles

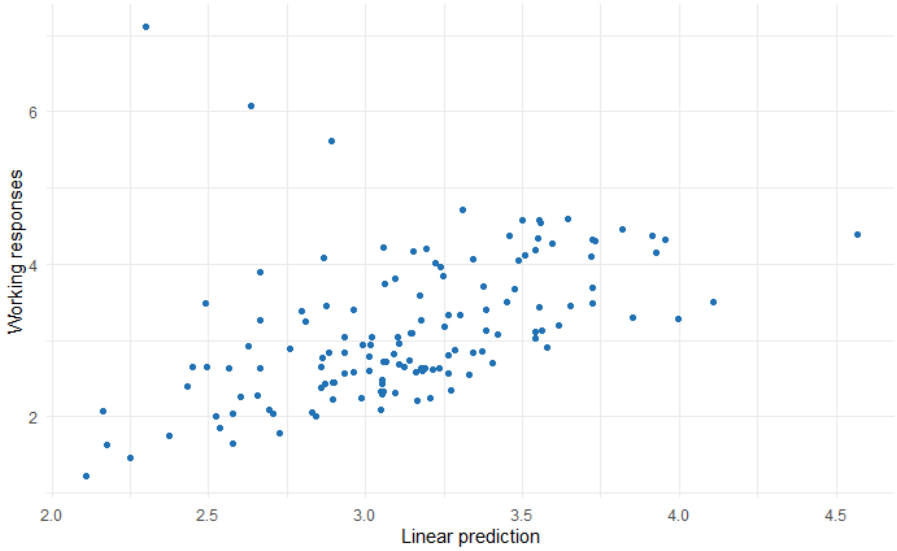

Figure 4.39: Final model's linear predictor vs working response

4.4.3 Log-normal scores - Identity link

Finally, the last scores can't use a straightforward GLM technique, since Log-normal distribution doesn't belong to the exponential family. The decision was to transform the scores by applying the log function, resulting in the new scores having a normal distribution and then proceed with a linear regression.

The stepwise algorithms, once more, reached the same model. This time, the proposed model has the variables (B1+Age+P1+S1+L1+Gender). Table 4.14 shows the estimates, standard errors and p-values for all these variables.

The estimates' signs go along with the previous models. Increasing B1, Age, L1 and S1's values lead to a higher severity score, while on the opposite side P1 and Gender=1

Table 4.14: Summary of the model obtained through the stepwise algorithm with selected variables' estimates, standard error and significance

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	1.405	0.369	3.809	0.000
B1	0.340	0.101	3.358	0.001
Age	0.026	0.006	4.187	0.000
L1	0.237	0.086	2.757	0.007
S1	0.240	0.104	2.312	0.022
P1	-0.169	0.087	-1.955	0.053
Gender1	-0.305	0.190	-1.607	0.110

(female) decrease the index. The significance of the variables is very reasonable, only Gender appears to add little value to the model (p-value=0.11). To check this hypothesis, an nested models comparison was performed between this model and the one removing Gender. The test's p-value was 0.11, which means that the larger model doesn't have significantly better results than the model without Gender. Therefore, the smaller model is chosen. The new estimates, errors and significances are presented in table 4.15.

The new model's estimates have small differences to the original stepwise model. And

Table 4.15: Summary of the final model with selected variables' estimates, standard error and significance

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	1.517	0.365	4.160	0.000
B1	0.317	0.101	3.139	0.002
Age	0.022	0.006	3.845	0.000
L1	0.255	0.086	2.963	0.004
S1	0.223	0.104	2.145	0.034
P1	-0.170	0.087	-1.956	0.053

the same can be said of the p-values. Only P1 has a value above 0.05, but it's really on the border. Another nested models comparison was executed between the model (B1+Age+L1+S1+P1) and this model withdrawing P1. The p-value was 0.05, again leaving doubts about the "right" decision. The model chosen was (B1+Age+L1+S1+P1) as, again, it has a PC of each group of variables.

The residual analysis starts with the plot of the deviance residuals against the fitted value, in figure 4.40. The plot shows the residuals gravitate around zero and the variance is approximately constant along the fitted values. The next plot, figure 4.41, shows the QQ plot of the Pearson residuals, and as can be seen the normal distribution fits very well this residuals. Moreover, when applying the KS test for normality, the p-value is 0.66, not rejecting the hypothesis of normality.

Finally, the plot that tests the suitability of the link function, figure 4.42, which in this case is just the fitted values against the real values, should show the points along a straight

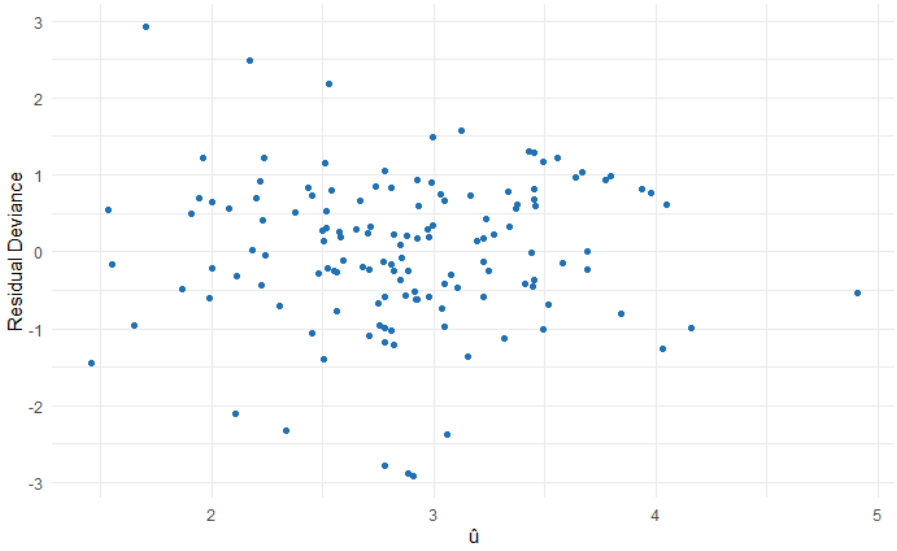

Figure 4.40: Final model's residual deviance vs fitted values

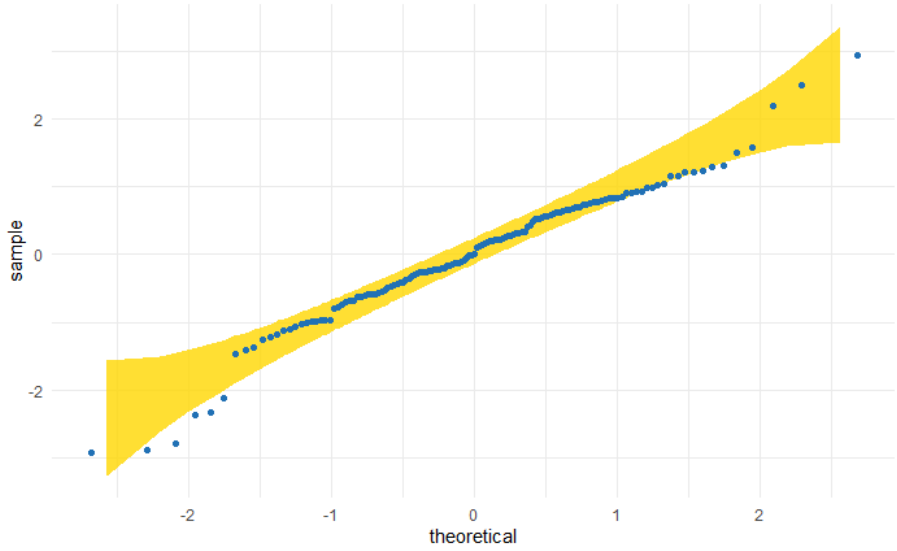

Figure 4.41: Pearson's residuals QQ plot compared to the normal distribution's quantiles

line. Again, an explicit line cannot be seen, but a linear trend seams to exist.

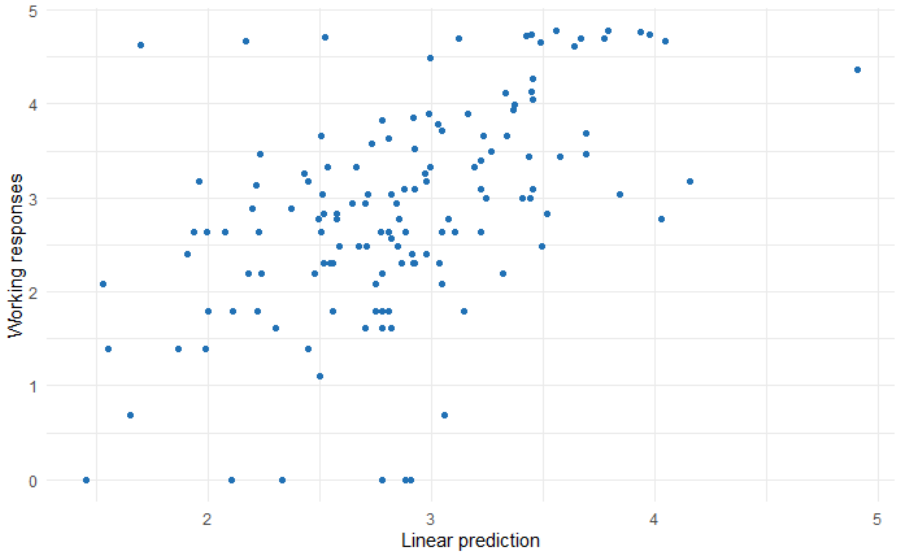

Figure 4.42: Final model's linear predictor vs working response

4.5 Models Quality

Initially, to assess the three models quality two ROC curves are plotted, one for the prediction of death and the other for predicting ICU hospitalization. After that, the ideal thresholds are calculated for each model with the aim of separating, first, the patients that died from the others, and, secondly, the ones in ICU from the patients in general hospitalization. In the end, the three models are compared to the original score used by the hospital.
Regarding the prediction of death, all three models clearly outperform the original Lung Ultrasound Score (LUS) (figure 4.43). All models have similar performances, but G-Inv tops the others with a AUC of 0.756 more than 1 tenth above LUS. In fact, the LUS death prediction confidence interval lower bound is always near the random classifier line.

After analysing these ROC curves, it's now possible to choose the ideal thresholds, for each model, that separates the patients that died, from those that didn't. Table 4.16 shows the ideal thresholds, corresponding to the point, in the ROC curves, closer to the ideal point (1,1).

The results are closely related to the ROC curves, and similar conclusions can be drawn. It's now possible to separate G-Log from LN as the latter has a better ideal threshold, because it leads to the same sensitivity but a higher specificity when compared to the former. G-Inv has the most balanced result, as the sensitivity and specificity are closer to each other than in the other models. The original score, LUS, has a high sensitivity, which means that it is good at identifying possible deaths, but, as it has a low specificity, there is a high volume of "false negatives", that can be translated as the score being high, although the patient hasn't a severe case.

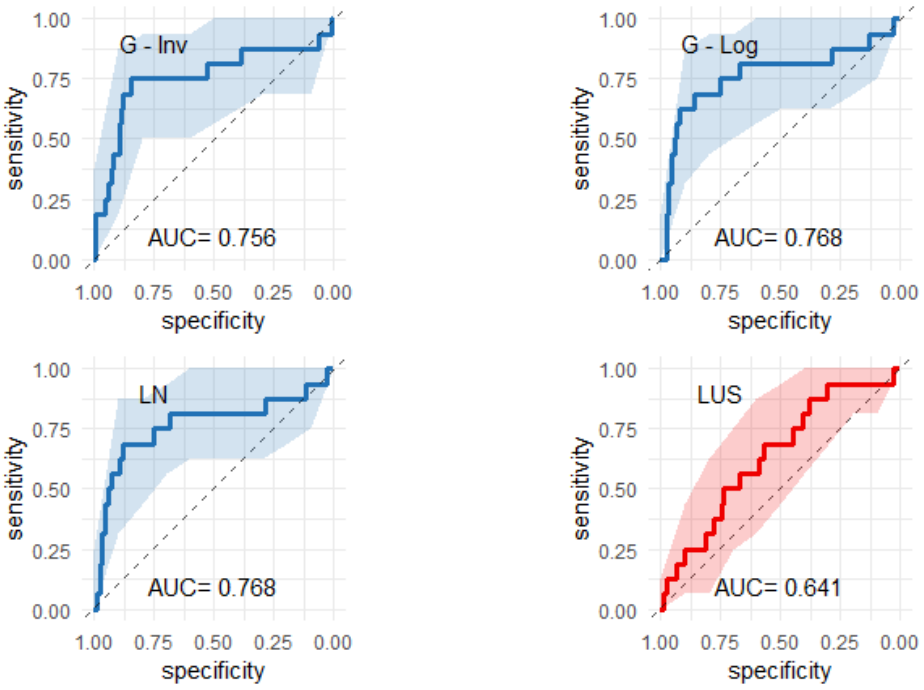

Figure 4.43: ROC curves of the three models compared to the original score (LUS) for death prediction with the respective confidence intervals obtained by bootstrap

Table 4.16: Ideal threshold for death prediction of the four models, obtained by the ROC curves' point closest to the ideal point (1,1), with the respective sensitivities and specificities

	G-Inv	G-Log	LN	LUS
Threshold	29.7	32.9	30.5	24.5
Sensitivity	75.0%	68.8%	68.8%	81.3%
Specificity	84.3%	86.0%	87.6%	50.4%

Considering now the prediction of ICU hospitalization at the time of the lung ultrasound (figure 4.44), the performances shift. In this case, the LUS is the best classifier, with the largest AUC of 0.875. The best of the three developed models is the LN model, with a AUC of 0.820, closely followed by the G-Log (0.818). In last place, is G-Inv.

Similarly to the death prediction, in the ICU case is also possible to reach the ideal thresholds for separating the patients in ICU from those in regular nursery hospitalization. This results are shown in table 4.17. It can be seen, that in this second evaluation, G-Log and LUS have the best results, the first one with the higher sensitivity and the second having the best specificity value. LN has only a slightly smaller performance when compared to G-Log, and G-Inv is clearly the worst in both metrics.

From this two evaluations, one can conclude that the models that best predict mortality,

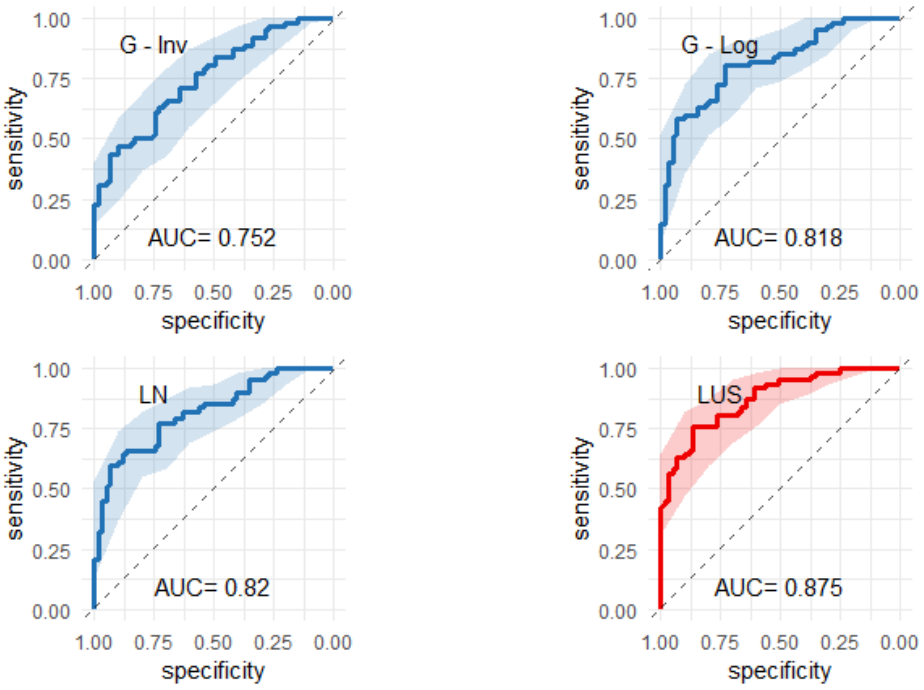

Figure 4.44: ROC curves of the three models compared to the original score (LUS) for ICU prediction with the respective confidence intervals obtained by bootstrap

Table 4.17: Ideal threshold for ICU hospitalization prediction of the four models, obtained by the ROC curves' point closest to the ideal point (1,1), with the respective sensitivities and specificities

	G-Inv	G-Log	LN	LUS
Threshold	21.9	21.2	16.1	27.5
Sensitivity	66.1%	80.6%	77.4%	74.2%
Specificity	69.5%	72.9%	72.9%	86.4%

have more trouble in ICU prediction and vice-versa. Another important remark is that the thresholds of the three proposed models are higher in death prediction than in ICU hospitalization prediction, which seems logical, as death is the worst outcome. However, LUS doesn't behave like the proposed models, and the thresholds are inverted, which reaffirms the handicap LUS in predicting death.

The most balanced model appears to be G-Log as, regarding the ideal thresholds, it outperforms the sensitivity and specificity of both G-Inv and LN in ICU prediction, and has a similar performance to these two in death prediction. LUS has a better performance after the threshold setting in ICU, but a very uneven performance in the death case (sensitivity is very good, 81.3%, but the specificity is low, 50.4%) which leads to a high number of false positives. The G-Log model can, therefore, be deemed as the most comprehensive model, as it performs well in all indicators.

5

Conclusion

This study had the main goal of creating an index, based on the information obtained through lung ultrasounds, aiming to find a good indicator of the diseases' severity on the patients, that could help the health workers make informed decisions. The hospital had already a score for ultrasounds, LUS, which was used as a baseline for the scores constructed. As severity is a subjective concept, three variables were used as proxy to severity: death, the need for ICU hospitalization and LOS. For this purpose, Generalized Linear Models was the methodology used as it is a very flexible regression technique.
To be able to make a connection between the ultrasound information and the severity through GLM's it is necessary to assign a probabilistic distribution to the response variable. As the point was for the score to be a good indicator of possibility of death, ICU hospitalization and LOS, the score was build using information of these three variables. Two distribution were found suitable for the scores, Gamma and Log-normal. Thus, two versions of the scores were used to develop models that were latter compared.
The information present on the lung ultrasounds of the dataset was divided in four parts for each lung in all individuals. Some of the variables, however, were co-linear which is problematic in the GLM's. For that reason, an ordinal Principal Components Analysis was performed for each variable, i.e., instead of having the information of four variables for each lung zone ($4 \times 8 = 32$), only the first two or three principal components associated to each variable were retained, leading to 10 new variables in total.
After the construction of the response variable and the improvement of the explanatory variables, the final models were obtained through stepwise algorithms and nested models comparison. Three final models were achieved: one using the Gamma distribution and the Inverse as link function (G-Inv); one also using Gamma distribution but with log as the link function (G-Log) and finally, using the Log-normal distribution, which were transformed by the log function, to have a normal distribution and a Gaussian regression was applied (LN). The number of variables selected by each model were different with G-Inv having only three variables in the final model (B1, Age and L1); G-Log and LN selected Age and all the variables' first principal component.
To assess the quality of the scores in predicting death and ICU hospitalization the study

of the respective ROC curves took place. All these results were also applied to the original LUS scores to better understand the gains of the new scores compared to the standard. In death prediction all three models had similar results and clearly outperformed LUS. Regarding the prediction of ICU hospitalization, at the time of the lung ultrasound, the original score has the better AUC, with LN and G-Log relatively close and G-Inv being the underperformer. The ideal thresholds have results that corroborate the AUC analysis and allow to better understand how to differentiate the patients.
The overall performance of the 4 scores leads to the conclusion that G-Log is the best balanced, having decent results in all analysis, and being the best when one analyses the ideal thresholds' sensitivity and specificity. Nevertheless, LUS has good evidences of being helpful, being the best in predicting ICU hospitalization.
The results hereby presented show that these mathematical approaches are valuable tools in the medical domain. Some improvements in future studies, however, can help solidify these results. The first would be to test these scores in a new, independent sample to see if the results are similar to the presented above. Furthermore, the application of this methodology to a bigger sample would also help as the larger the sample the more realistic are the proportions of death and ICU hospitalizations. Thus, would also lead to a more correct assessment on the distribution of LOS. The contribution that statistical indicators have on medical environments is increasing and can only have a positive impact as these tools are not to substitute the medical analysis but to be one more component that can be used by the decision makers.

Bibliography

[1] G. Li et al. "Coronavirus infections and immune responses". In: *Journal of Medical Virology* 92.4 (2020), pp. 424–432. ISSN: 10969071. DOI: 10.1002/jmv.25685 (cit. on p. 3).

[2] J. M. Miller et al. "A Guide to Utilization of the Microbiology Laboratory for Diagnosis of Infectious Diseases: 2018 Update by the Infectious Diseases Society of America and the American Society for Microbiology". In: *Clinical Infectious Diseases* 67.6 (2018), e1–e94. ISSN: 15376591. DOI: 10.1093/cid/ciy381 (cit. on p. 3).

[3] CDC. *Nucleic Acid Amplification Tests (NAATs)*. 2021. URL: https://www.cdc.gov/coronavirus/2019-ncov/lab/naats.html. (accessed: 19.04.2022) (cit. on p. 3).

[4] J. Penney et al. "Cycle threshold values and SARS-CoV-2: Relationship to demographic characteristics and disease severity". In: *Journal of Medical Virology* (2022). ISSN: 0146-6615. DOI: 10.1002/jmv.27752. URL: https://onlinelibrary.wiley.com/doi/10.1002/jmv.27752 (cit. on p. 3).

[5] R. A. Lee et al. "Performance of Saliva, Oropharyngeal Swabs, and Nasal Swabs for SARS-CoV-2 Molecular Detection: a Systematic Review and Meta-analysis". In: *Journal of Clinical Microbiology* 59.5 (2021), e02881–20. DOI: 10.1128/JCM.02881-20. eprint: https://journals.asm.org/doi/pdf/10.1128/JCM.02881-20. URL: https://journals.asm.org/doi/abs/10.1128/JCM.02881-20 (cit. on p. 3).

[6] B. Bruzzone et al. "Comparative diagnostic performance of rapid antigen detection tests for COVID-19 in a hospital setting". In: *International Journal of Infectious Diseases* 107 (2021), pp. 215–218. ISSN: 18783511. DOI: 10.1016/j.ijid.2021.04.072. URL: https://doi.org/10.1016/j.ijid.2021.04.072 (cit. on p. 4).

[7] O. O. Abayomi-Alli et al. "An Ensemble Learning Model for COVID-19 Detection from Blood Test Samples". In: *Sensors* 22.6 (2022). ISSN: 14248220. DOI: 10.3390/s22062224 (cit. on p. 4).

[8] E. Ingberg et al. "RT-PCR cycle threshold value in combination with visual scoring of chest computed tomography at hospital admission predicts outcome in COVID-19". In: *Infectious Diseases* 54.6 (2022), pp. 431–440. ISSN: 23744235. DOI: 10.1080/23744235.2022.2035428. URL: https://doi.org/10.1080/23744235.2022.2035428 (cit. on p. 4).

[9] C. K. Kwoh, K. H. Lee, and T. V. Le. "Using survival models to analyze the effects of social attributes on length of stay of stroke patients". In: *2nd International Conference on Biomedical and Pharmaceutical Engineering, ICBPE 2009 - Conference Proceedings* 4 (2009), pp. 1–5. DOI: 10.1109/ICBPE.2009.5384062 (cit. on p. 5).

[10] R. E. Al Mamlook, H. F. Bzizi, and S. Chen. "Evaluate Performance Risk Score in Patients Suffering from Lung Cancer Using Survival Analysis of Statistics". In: *IEEE International Conference on Electro Information Technology* 2020-July (2020), pp. 145–150. ISSN: 21540373. DOI: 10.1109/EIT48999.2020.9208342 (cit. on p. 5).

[11] R. Yamamoto et al. "Modified abbreviated burn severity index as a predictor of in-hospital mortality in patients with inhalation injury: development and validation using independent cohorts". In: *Surgery Today* 51.2 (2021), pp. 242–249. ISSN: 14362813. DOI: 10.1007/s00595-020-02085-5 (cit. on p. 5).

[12] E. Stühler et al. "Framework for personalized prediction of treatment response in relapsing remitting multiple sclerosis". In: *BMC Medical Research Methodology* 20.1 (2020), pp. 1–16. ISSN: 14712288. DOI: 10.1186/s12874-020-0906-6 (cit. on p. 5).

[13] J. D. Schoufour et al. "Design of a frailty index among community living middle-aged and older people: The Rotterdam study". In: *Maturitas* 97 (2017), pp. 14–20. ISSN: 18734111. DOI: 10.1016/j.maturitas.2016.12.002. URL: http://dx.doi.org/10.1016/j.maturitas.2016.12.002 (cit. on p. 6).

[14] E. Bernabé et al. "The T-Health index: a composite indicator of dental health". In: *European Journal of Oral Sciences* 117.4 (2009), pp. 385–389. ISSN: 09098836. DOI: 10.1111/j.1600-0722.2009.00649.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1600-0722.2009.00649.x (cit. on p. 6).

[15] J. L. Miller et al. "Prediction models for severe manifestations and mortality due to <scp>COVID</scp> -19: A systematic review". In: *Academic Emergency Medicine* 29.2 (2022), pp. 206–216. ISSN: 1069-6563. DOI: 10.1111/acem.14447. URL: https://onlinelibrary.wiley.com/doi/10.1111/acem.14447 (cit. on p. 7).

[16] S. Bolourani et al. "A machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19: Model development and validation". In: *Journal of Medical Internet Research* 23.2 (2021), pp. 1–15. ISSN: 14388871. DOI: 10.2196/24246 (cit. on p. 7).

[17] Z. Chen et al. "A risk score based on baseline risk factors for predicting mortality in COVID-19 patients". In: *Current Medical Research and Opinion* 37.6 (2021), pp. 917–927. ISSN: 14734877. DOI: 10.1080/03007995.2021.1904862. URL: https://doi.org/10.1080/03007995.2021.1904862 (cit. on p. 7).

[18] A. K. Das, S. Mishra, and S. S. Gopalan. "Predicting CoVID-19 community mortality risk using machine learning and development of an online prognostic tool". In: *PeerJ* 8 (2020), pp. 1–12. ISSN: 21678359. DOI: 10.7717/peerj.10083 (cit. on p. 7).

[19] S. Goodacre et al. "Derivation and validation of a clinical severity score for acutely ill adults with suspected COVID-19: The PRIEST observational cohort study". In: *PLoS ONE* 16.1 January (2021), pp. 1–19. ISSN: 19326203. DOI: 10.1371/journal.pone.0245840 (cit. on p. 8).

[20] N. Xirouchaki et al. "Lung ultrasound in critically ill patients: Comparison with bedside chest radiography". In: *Intensive Care Medicine* 37.9 (2011), pp. 1488–1493. ISSN: 03424642. DOI: 10.1007/s00134-011-2317-y (cit. on p. 9).

[21] Q. Deng et al. "Semiquantitative lung ultrasound scores in the evaluation and follow-up of critically ill patients with COVID-19: a single-center study". In: *Academic Radiology* 27.10 (2020), pp. 1363–1372. ISSN: 18784046. DOI: 10.1016/j.acra.2020.07.002. URL: https://doi.org/10.1016/j.acra.2020.07.002 (cit. on pp. 9, 10).

[22] G. Volpicelli et al. "International evidence-based recommendations for point-of-care lung ultrasound". In: *Intensive Care Medicine* 38.4 (2012), pp. 577–591. ISSN: 03424642. DOI: 10.1007/s00134-012-2513-4 (cit. on p. 9).

[23] J. C. G. de Alencar et al. "Lung ultrasound score predicts outcomes in COVID-19 patients admitted to the emergency department". In: *Annals of Intensive Care* 11.1 (2021). ISSN: 21105820. DOI: 10.1186/s13613-020-00799-w (cit. on pp. 9, 10).

[24] D. M. Tierney et al. "Pulmonary ultrasound scoring system for intubated critically ill patients and its association with clinical metrics and mortality: A prospective cohort study". In: *Journal of Clinical Ultrasound* 46.1 (2018), pp. 14–22. ISSN: 10970096. DOI: 10.1002/jcu.22526 (cit. on pp. 9, 11).

[25] G. Soldati et al. "Proposal for International Standardization of the Use of Lung Ultrasound for Patients With COVID-19". In: *Journal of Ultrasound in Medicine* 39.7 (2020), pp. 1413–1419. ISSN: 15509613. DOI: 10.1002/jum.15285 (cit. on pp. 9–11).

[26] P. Trias-Sabrià et al. "Lung ultrasound score to predict outcomes in COVID-19". In: *Respiratory Care* 66.8 (2021), pp. 1263–1270. ISSN: 19433654. DOI: 10.4187/RESPCARE.08648 (cit. on p. 10).

[27] C. Baloescu et al. "Automated Lung Ultrasound B-Line Assessment Using a Deep Learning Algorithm". In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 67.11 (2020), pp. 2312–2320. ISSN: 15258955. DOI: 10.1109/TUFFC.2020.3002249 (cit. on p. 11).

[28] S. Bagon et al. "Integrating domain knowledge into deep networks for lung ultrasound with applications to covid-19". In: *IEEE Transactions on Medical Imaging* 41.3 (2022), pp. 571–581. ISSN: 1558254X. DOI: 10.1109/TMI.2021.3117246 (cit. on p. 11).

[29] J. Chen et al. "Quantitative Analysis and Automated Lung Ultrasound Scoring for Evaluating COVID-19 Pneumonia with Neural Networks". In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 68.7 (2021), pp. 2507–2515. ISSN: 15258955. DOI: 10.1109/TUFFC.2021.3070696 (cit. on p. 12).

[30] W. J. Conover. *Practical Nonparametric Statistics*. Third. John Wiley & Sons, 1999, p. 584. ISBN: 0471160687 (cit. on pp. 15, 16).

[31] J. A. Nelder and R. W. M. Wedderburn. "Generalized Linear Models". In: *Journal of the Royal Statistical Society. Series A (General)* 135.3 (1972), p. 370. ISSN: 00359238. DOI: 10.2307/2344614. URL: https://www.jstor.org/stable/10.2307/2344614?origin=crossref (cit. on p. 17).

[32] M. A. A. Turkman and G. L. Silva. "Modelos Lineares Generalizados-da teoria à prática". In: *Sociedade Portuguesa de Estatística, Lisboa* (2000) (cit. on pp. 17, 70).

[33] G. Casella and R. L. Berger. *Statistical inference*. Cengage Learning, 2021 (cit. on p. 20).

[34] P. Mair. "Gifi Methods". In: *Modern Psychometrics with R*. Cham: Springer International Publishing, 2018, pp. 231–256. ISBN: 978-3-319-93177-7. DOI: 10.1007/978-3-319-93177-7_8. URL: https://doi.org/10.1007/978-3-319-93177-7_8 (cit. on p. 33).

[35] G. Michailidis and J. de Leeuw. "The Gifi system of descriptive multivariate analysis". In: *Statistical Science* 13.4 (1998). ISSN: 0883-4237. DOI: 10.1214/ss/1028905828. URL: https://projecteuclid.org/journals/statistical-science/volume-13/issue-4/The-Gifi-system-of-descriptive-multivariate-analysis/10.1214/ss/1028905828.full (cit. on p. 33).

[36] L. Gonçalves et al. "ROC Curve Estimation: An Overview". In: *REVSTAT-Statistical Journal* 12.1 (2014), 1–20. DOI: 10.57805/revstat.v12i1.141. URL: https://revstat.ine.pt/index.php/REVSTAT/article/view/141 (cit. on p. 36).

[37] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022. URL: https://www.R-project.org/ (cit. on p. 37).

[38] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC. Boston, MA, 2020. URL: http://www.rstudio.com/ (cit. on p. 37).

[39] T. Sing et al. "ROCR: visualizing classifier performance in R". In: *Bioinformatics* 21.20 (2005), p. 7881. URL: http://rocr.bioinf.mpi-sb.mpg.de (cit. on p. 37).

[40] X. Robin et al. "pROC: an open-source package for R and S+ to analyze and compare ROC curves". In: *BMC Bioinformatics* 12 (2011), p. 77 (cit. on p. 37).

[41] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org (cit. on p. 37).

[42] E. Neuwirth. *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-3. 2022. URL: https://CRAN.R-project.org/package=RColorBrewer (cit. on p. 37).

[43] A. Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.4.0. 2020. URL: https://CRAN.R-project.org/package=ggpubr (cit. on p. 37).

[44] A. Almeida, A. Loy, and H. Hofmann. *ggplot2 Compatible Quantile-Quantile Plots in R*. 2. 2018, pp. 248–261. URL: https://doi.org/10.32614/RJ-2018-051 (cit. on p. 37).