Pontifícia Universidade Católica
do Rio de Janeiro

Igor Tona Peres

Essays on length of stay prediction in Intensive Care Units

Tese de Doutorado

Thesis presented to the Programa de Pós-Graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia de Produção.

Advisor : Prof. Fernando Luiz Cyrino Oliveira
Co-adviser: Prof. Silvio Hamacher

Rio de Janeiro
May 2021

Pontifícia Universidade Católica
do Rio de Janeiro

Igor Tona Peres

Essays on length of stay prediction in Intensive Care Units

Thesis presented to the Programa de Pós-Graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia de Produção. Approved by the Examination Committee:

Prof. Fernando Luiz Cyrino Oliveira
Advisor
Departmento de Engenharia Industrial PUC-Rio

Prof. Silvio Hamacher
Co-adviser
Departamento de Engenharia Industrial PUC-Rio

Prof. Fernando Augusto Bozza
FIOCRUZ

Prof. Jorge Ibrain Figueira Salluh
Instituto D'Or de Pesquisa e Ensino

Prof. Fernanda Araújo Baião Amorim
Departamento de Engenharia Industrial - PUC-RIO

Prof. Benjamin Dalmas
EMSE

Rio de Janeiro, May 7th, 2021

Igor Tona Peres

Igor Peres holds a Bachelor of Science degree in Industrial Engineering from Federal University of Rio de Janeiro, Brazil (2010) and a Master's degree in Industrial Engineering from Pontifical Catholic University of Rio de Janeiro (PUC-Rio) (2017). He was project manager at Trilha da Inovação from 2010 to 2014. Since 2018, he has worked at the Institute of Technical-Scientific Software Development of PUC-Rio (Tecgraf / PUC-Rio) as researcher fellow. Has experience in Industrial Engineering, with emphasis on Production Management. He works in the areas of Data Science, Machine Learning, Operations Research, Simulation, Optimization, and Health Management.

Bibliographic data

Peres, Igor Tona

Essays on length of stay prediction in Intensive Care Units / Igor Tona Peres; advisor: Fernando Luiz Cyrino Oliveira; co-adviser: Silvio Hamacher. – 2021.
124 f. : il. color. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Industrial, 2021.
Inclui bibliografia

1. Engenharia Industrial – Teses. 2. Ciência de dados. 3. Aprendizado de máquina. 4. Modelos preditivos. 5. Tempo de permanência. 6. Unidades de terapia intensiva. I. Oliveira, Fernando Luiz Cyrino. II. Hamacher, Silvio. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Industrial. IV. Título.

CDD: 658.5

To God, to my parents, Fabíola and Sergio,
and to my brothers Ian and Gabriel.

Acknowledgments

To my parents, Sergio and Fabíola, I thank you for all your support, education, affection, love, and trust. Without them none of this would have been possible.
To my brothers, Gabriel and Ian, for the support at all times that I needed, and to all my family for the values acquired.
To my advisors Fernando Cyrino and Silvio Hamacher, for all support, attention, corrections, incentives and opportunities during the development of the thesis. Thank you for all the teachings.
To the physicians Fernando Bozza and Jorge Salluh, for the great support in clinical analyses.
To DEI / PUC-Rio professors for all their shared knowledge and learning opportunities.
To my colleagues and friends at PUC-Rio, thank you for sharing your doubts, experiences and for making this whole journey a little easier.
To CNPQ, CAPES and PUC-Rio, for the aid granted and for the excellent environment, materials and professors available, without whom this work would not have been possible.
To all those who, in some way, contributed to this work, my thanks.
This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.
This study was financed in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ).

Abstract

Peres, Igor Tona; Oliveira, Fernando Luiz Cyrino (Advisor); Hamacher, Silvio (Co-Advisor). **Essays on length of stay prediction in Intensive Care Units**. Rio de Janeiro, 2021. 124p. Tese de Doutorado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

The length of stay (LoS) in Intensive Care Units (ICU) is one of the most used metrics for resource use. This thesis proposes a structured data-driven methodology to approach three main demands of ICU managers. First, we propose a model to predict the individual ICU length of stay, which can be used to plan the number of beds and staff required. Second, we develop a model to predict the risk of prolonged stay, which helps identifying prolonged stay patients to drive quality improvement actions. Finally, we build a case-mix-adjusted efficiency measure (SLOSR) capable of performing non-biased benchmarking analyses between ICUs. To achieve these objectives, we divided the thesis into the following specific goals: (i) to perform a literature review and meta-analysis of factors that predict patient's LoS in ICUs; (ii) to propose a data-driven methodology to predict the numeric ICU LoS and the risk of prolonged stay; and (iii) to apply this methodology in the context of a big set of ICUs from mixed-type hospitals. The literature review results presented the main risk factors that should be considered in future prediction models. Regarding the predictive model, we applied and validated our proposed methodology to a dataset of 109 ICUs from 38 different Brazilian hospitals. The included dataset contained a total of 99,492 independent admissions from January 01 to December 31, 2019. The predictive models to numeric ICU LoS and to the risk of prolonged stay built using our data-driven methodology presented accurate results compared to the literature. The proposed models have the potential to improve the planning of resources and early identifying prolonged stay patients to drive quality improvement actions. Moreover, we used our prediction model to build a non-biased measure for ICU benchmarking, which was also validated in our dataset. Therefore, this thesis proposed a structured data-driven guide to generating predictions to ICU LoS adjusted to the specific environment analyzed.

Keywords

Data science; Machine learning; Predictive modeling; Length of stay; Intensive care units.

Resumo

Peres, Igor Tona; Oliveira, Fernando Luiz Cyrino; Hamacher, Silvio. **Ensaios em predição do tempo de permanência em Unidades de Terapia Intensiva**. Rio de Janeiro, 2021. 124p. Tese de Doutorado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

O tempo de permanência (LoS) é uma das métricas mais utilizadas para avaliar o uso de recursos em Unidades de Terapia Intensiva (UTI). Esta tese propõe uma metodologia estruturada baseada em dados para abordar três principais demandas de gestores de UTI. Primeiramente, será proposto um modelo de predição individual do LoS em UTI, que pode ser utilizado para o planejamento dos recursos necessários. Em segundo lugar, tem-se como objetivo desenvolver um modelo para predizer o risco de permanência prolongada, o que auxilia na identificação deste tipo de paciente e assim uma ação mais rápida de intervenção no mesmo. Finalmente, será proposto uma medida de eficiência ajustada por "case-mix" capaz de realizar análises comparativas de "benchmark" entre UTIs. Os objetivos específicos são: (i) realizar uma revisão da literatura dos fatores que predizem o LoS em UTI; (ii) propor uma metodologia data-driven para predizer o LoS individual do paciente na UTI e o seu risco de longa permanência; e (iii) aplicar essa metodologia no contexto de um grande conjunto de UTIs de diferentes tipos de hospitais. Os resultados da revisão da literatura apresentaram os principais fatores de risco que devem ser considerados em modelos de predição. Em relação ao modelo preditivo, a metodologia proposta foi aplicada e validada em um conjunto de dados de 109 UTIs de 38 diferentes hospitais brasileiros. Este conjunto continha um total de 99.492 internações de 01 de janeiro a 31 de dezembro de 2019. Os modelos preditivos construídos usando a metodologia proposta apresentaram resultados precisos comparados com a literatura. Estes modelos propostos têm o potencial de melhorar o planejamento de recursos e identificar precocemente pacientes com permanência prolongada para direcionar ações de melhoria. Além disso, foi utilizado o modelo de predição proposto para construir uma medida não tendenciosa para benchmarking de UTIs, que também foi validada no conjunto de dados estudado. Portanto, esta tese propôs um guia estruturado baseado em dados para gerar predições para o tempo de permanência em UTI ajustadas ao contexto em que se deseja avaliar.

Palavras-chave

Ciência de dados; Aprendizado de máquina; Modelos preditivos; Tempo de permanência; Unidades de terapia intensiva.

Table of contents

List of figures

List of tables

1
Introduction

Hospitals are facing continuous pressure to improve efficiency and reduce costs. Intensive Care Units (ICUs) are complex environments that provide expensive care [Halpern and Pastores, 2015; Halpern et al., 2004]. Despite differences in definitions of prolonged length of stay (LoS), studies have shown that a small percentage of ICU patients (4% to 11%) presented a prolonged LoS [Arabi et al., 2002; Higgins et al., 2003b; Laupland et al., 2006; Schoffelen et al., 2010; Zampieri et al., 2014; Zimmerman et al., 2006]. However, those few patients account for a large proportion of ICU days (40% to 52%) [Arabi et al., 2002; Higgins et al., 2003b; Laupland et al., 2006; Schoffelen et al., 2010; Zampieri et al., 2014; Zimmerman et al., 2006]. Since hospital costs are strongly related to ICU LoS [Arabi et al., 2002; Halpern et al., 2004; Higgins et al., 2003b; Kahn et al., 2008; Laupland et al., 2006; Schoffelen et al., 2010; Zampieri et al., 2014; Zimmerman et al., 2006], few ICU admissions must account for a large proportion of hospital costs. Therefore, the early identification of prolonged stay patients can assist in improving unit efficiency. In short, the main reasons for hospital administrators to predict ICU LoS are threefold: (i) planning the number of ICU resources required; (ii) identifying patients with greater risk of prolonged stay aiming to drive quality improvement actions; and (iii) enabling case-mix[1] adjustments for benchmarking analysis [Kahn et al., 2008; Marik and Hedman, 2000; Verburg et al., 2017; Zampieri et al., 2014].

This thesis developed a structured data-driven methodology to deal with each of these three hospital administrators' demands. First, we propose a model to predict the individual ICU length of stay, which can be used to plan the number of beds and staff required. Second, we develop a model to predict the risk of prolonged stay, which helps identifying prolonged stay patients to drive immediate quality improvement. Finally, we build a case-mix-adjusted efficiency measure capable of performing non-biased benchmarking analyses between ICUs. To achieve these objectives, we divided the thesis into the following specific goals: (i) perform a systematic literature review (SLR) and meta-analysis of factors that predict patient's LoS in ICUs; (ii) propose a data-

[1]Case-mix: The relative numbers of various types of patients being treated as categorized by disease-related groups, severity of illness, rate of consumption of resources, and other indicators; used as a tool for managing and planning health care services.

driven methodology to predict the numeric ICU LoS and the risk of prolonged stay; and (iii) apply this methodology in the context of a big set of ICUs from mixed-type hospitals. The literature review results presented the main risk factors that should be considered in future prediction models. Regarding the predictive model, we applied the methodology to a dataset of 109 ICUs from 38 different Brazilian hospitals.

This thesis is organized into five chapters, as follows. Chapter 2 will present an SLR and meta-analysis of factors that predict the patient length of stay in the ICU. Chapter 3 will present a data-driven methodology to predict the numeric ICU LoS and the risk of prolonged stay, and a methodology to perform benchmarking analysis between ICUs. Chapter 4 presents an application in a dataset with 109 mixed-type ICUs from 38 different Brazilian hospitals. Chapter 5 presents the conclusions and future perspectives.

2
What factors predict length of stay in the Intensive Care Unit? Systematic Review and Meta-Analysis

This chapter aims to analyze what factors were associated with ICU length of stay. To accomplish this goal, we performed a systematic review and meta-analysis of papers that reported risk factors for ICU LoS. This study was published in the Journal of Critical Care [Peres et al., 2020].

2.1 Introduction

Recent studies have analyzed the association of factors with ICU LoS. A variety of designs have been used, including different factors related to prolonged stay and reporting a range of statistical measures (i.e., correlations, t-test statistics, mean, standard deviation). The existing reviews deal with specific populations – Neonatal Unit [Seaton et al., 2016] and after coronary artery bypass grafting ICU [Atashi et al., 2018] patients – or focus on models used to predict LoS [Atashi et al., 2018; Awad et al., 2017; Verburg et al., 2017]. Although a comprehensive synthesis of factors associated with ICU LoS that considers general adult population would be of great interest to researchers and administrators, to the best of our knowledge, no updated Systematic Literature Review (SLR) or meta-analysis exists in this field.

In the present study, we performed an SLR and meta-analysis to understand the risk factors of ICU LoS. The goals were twofold: to provide the characteristics of existing studies and to perform a meta-analysis of the statistics reported.

2.2 Materials and methods

The SLR and meta-analysis were conducted and reported according to the recommendations of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [Liberati et al., 2009; Moher et al., 2009]. The protocol was registered on PROSPERO (ID: CRD42019121642) in April 2019. In what follows, we describe our search strategy, recount eligibility

criteria for study selection, and how data extraction, quality assessment, and statistical analysis were performed.

2.2.1 Information Source and Search Strategy

From inception to November 20, 2018, we searched MEDLINE, Embase and Scopus databases. The searching process was limited to the English language and the publication types "article"; "article in press"; and "review". The search comprised the fields "title"; "abstracts"; and "keywords" and no restriction was made for the publication period. We used the following query: ("ICU" or "Intensive Care" or "Critical Care" or "Critically Ill") and ("length of stay") and ("predict" or "predictive" or "prediction" or "predictor" or "prognosis" or "prognostic"). Moreover, an advanced search was performed to retrieve different spelling occurrences of the keywords (e.g., "length-of-stay" instead of "length of stay") in both the singular and plural forms. The review was subsequently updated until May 22, 2020, with a forward snowballing search.

2.2.2 Study selection

The study selection was fourfold: (i) formulating eligibility criteria; (ii) abstract reading and selection for full-text reading; (iii) full-text reading and selection for SLR and meta-analysis; (iv) including new studies by backward and forward search [Liberati et al., 2009; Moher et al., 2009; Thomé et al., 2016].

We considered the following eligibility criteria for study inclusion in the SLR: (i) study deals with general adult ICU population; (ii) study analyzes one or more factors related to patient LoS in the ICU; (iii) study does not deal strictly with research in the medical field related to clinical treatment or diseases; (iv) study does not deal with experiments in animals; and (v) systematic reviews. For meta-analysis inclusion, we also consider the following criteria: (vi) studies reporting appropriate statistics (i.e., correlations, t-test statistics, mean, standard deviation) that could be converted to one of the following effect sizes: partial correlation and standardized mean difference. The studies whose abstract or full-text did not meet any of the above criteria were excluded. Inclusion and exclusion criteria were pilot tested randomly in ten papers, and disagreements were debated by at least three authors until an agreement about exclusion was reached. Studies that were not available for online download were also excluded from the analysis. The previous systematic

reviews related to the research topic were used to enlighten the discussion section only and to enrich the comparisons with previous results.

2.2.3 Data Extraction

We developed a data extraction sheet, pilot-test it on ten randomly-selected included studies, and refined it accordingly. The following data were recorded (when available): study characteristics (cohort size, year, continent, setting, and design), patient characteristics (inclusion and exclusion criteria), statistical method used, type of LoS measured (continuous or categorical), basic statistics of ICU LoS (e.g., mean and standard deviation), time of variables' measurement, significant and non-significant variables, and statistics presented (e.g., correlations, t-test statistics).

2.2.4 Quality Assessment

Study quality was discussed using an adaptation of the Quality In Prognostic Studies (QUIPS) tool [Hayden et al., 2013]. We considered all domains of QUIPS tool: study participation (a cohort should include medical and surgical patients consecutively admitted to the ICU and not a specific population); study attrition (the sample available for analysis should be representative); factors measurement (the definition and measurement of factors and the methods used for missing data should be appropriate); outcome measurement (the definition and measurement of ICU LoS should be proper); study confounding (the model should be adjusted by potential confounders) and statistical analysis and reporting (the model building should be appropriate, considering validation and with no selective reporting of results). We included a rate for each article based on the attendance of the domains. This rate could vary from one to six, according to the number of domains attended. The total number of domains attended in each paper was called "number of stars".

2.2.5 Statistical analysis

A meta-analysis was performed, when possible, seeking to estimate the average relationship across studies between a given factor and the ICU LoS. We used the following effect sizes: (1) the partial correlation rp [Aloe, 2014; Aloe and Thompson, 2013] for studies reporting multiple regression statistics; and (2) the standardized mean difference (Cohen's d) [Cohen, 1988] for studies reporting means of ICU LoS for each factor category. Where data were reported

as medians and interquartile ranges, it was converted to means and standard deviation [Luo et al., 2018; Wan et al., 2014]. The I^2 test was used to describe the proportion of the total variation in the study estimates that is due to heterogeneity in the meta-analysis [Higgins et al., 2003a]. The effect sizes, as well as the associated 95% confidence intervals (CI), were computed with a random-effects model due to expected heterogeneity. Publication bias was detected by Begg's test [Begg and Mazumdar, 1994]. Statistical analyses were performed using the package Metafor in R software version 3.4 [R Core Team, 2018]. To assess the robustness of the estimates' stability, a sensitivity analysis was also performed.

2.3 Results

2.3.1 Study selection

The initial search identified 8935 papers. After removing duplicates, 6906 articles were screened for eligibility based on their title or abstract. The remaining 168 studies were screened based on their complete text, leaving a total of 85 research articles. After screening references, 16 papers were included, leaving a total of 101 research articles. The review was updated until May 2020 with a forward search on articles that cited these 101 papers, resulting in the addition of 12 new articles. We presented the complete searching process in Figure 2.1.

2.3.2 Summary of studies

Only nine studies used a cohort size smaller than 100 patients [Brascher et al., 2020; Cander et al., 2011; da Silva et al., 2015; Ely et al., 2001; Limaye et al., 2011; Makrygiannis et al., 2018; Piva et al., 2015; Thorevska et al., 2003; Zafar et al., 2014], and the median cohort, first and third quartile were, respectively, 443, 200, and 2,588 patients. The total cohort of all papers was equal to 2,163,424 patients. Most studies consider for inclusion criteria patients with ICU LoS longer than 24h and with age older than 18 years old. There was a range of exclusion criteria used in the studies, as follows: presence of missing values, burns, surgical patients, post-operative patients, pregnant women, patients not expected to survive, readmissions, admissions from or transfers to another ICU, and other less common exclusions.

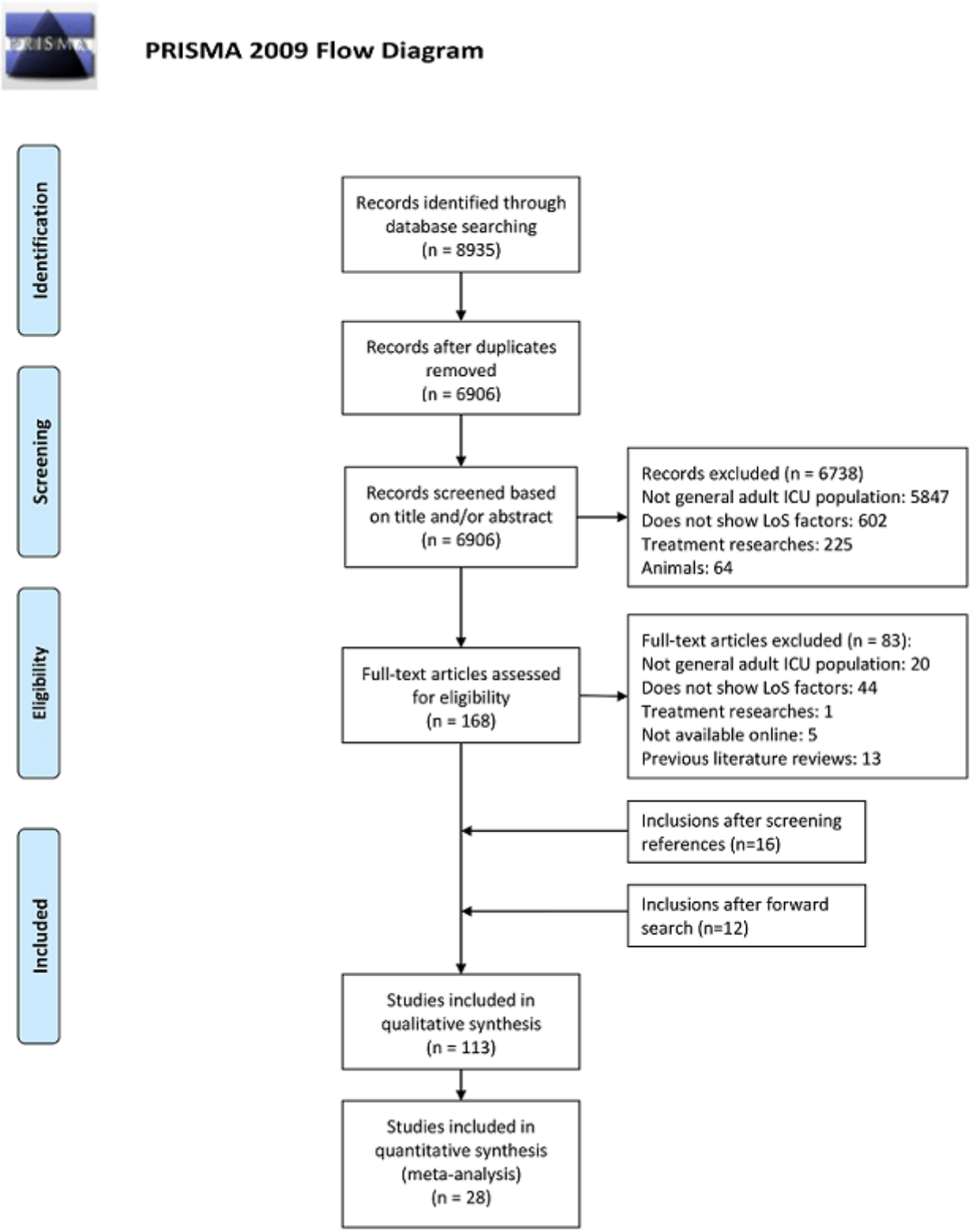


Figure 2.1: Flow diagram of study inclusion.

Regarding the time of variables' measurement, over half of the studies took the measurements on admission or within the first 24 h. The large majority of articles performed some type of univariate analysis. Some articles also used multivariate methods in order to adjust the effects by potential confounders. Most studies consider the ICU LoS as a continuous dependent variable, and few articles consider the LoS as a categorical variable (i.e., prolonged stay or normal stay). Of these, the majority considered as prolonged ICU stay admissions greater than 14 days [Arabi et al., 2002; Higgins et al., 2003b; Laupland et al., 2006; Schoffelen et al., 2008; Zampieri et al., 2014].

2.3.3 Quality Assessment

The overall quality of the studies was moderate: 69 articles (61%) met at least five of the six QUIPS domains analyzed, and 101 articles (89%) attended at least four domains. The number of articles that attended each quality domain is as follows: 84 for study participation, 108 for study attrition, 93 for factors measurement, 112 for outcome measurement, 38 for study confounding, and 90 for statistical analysis and reporting.

2.3.4 Risk factors of ICU stay

The 113 identified studies presented a total of 163 possible risk factors of ICU LoS. Because of this great number, we selected for discussion only factors analyzed by more than one article. Therefore, 89 factors were analyzed and will be discussed next. These variables were grouped into broad categories of patient demographics; severity scores; characteristics of admission; interventions; clinical conditions; acute diagnoses; APACHE IV diagnoses; chronic health items; reasons for ICU admission; and clinical information. Table 2.1 summarizes our findings related to predictors of ICU stay and provides all factors assessed at least three times. The table presents the number of studies that found each factor to be significant (or non-significant), the relative percentage "N (%)", the corresponding references and also the references when the effect was positive or negative.

We begin by reporting on factors related to the patient demographics, namely age, gender, and body mass index (BMI). Age was the most analyzed factor among all articles, but the results were not conclusive. Most studies that found age to be significant considered this factor as a categorical variable, and noted that: from 18 to about 60 years, the greater the age, the greater the ICU LoS; and from 61 years, the higher the age, the shorter the ICU LoS

Factors	Signif. N	Signif. (%)	References +	References -	References All*	Non-signif. N	Non-signif. (%)	References All*
Patient demographics								
Age	10	40%	[37,51,58]	[42]	[7,37,39,42,50,51,58,64–66]	15	60%	[3,4,6,8,28,41,47,49,52,54,58,59,61,62,67]
Gender	3	18%	*	*	[39,42,58]	14	82%	[3,6–8,28,37,51,52,58,61,64,65]
Body mass index (BMI)	4	57%	[52,68]	*	[44,52,68,69]	3	43%	[3,70,71]
Severity scores								
APACHE II	8	53%	[6,11,41,48,72,73]	*	[4,6,8,11,41,48,72,73]	7	47%	[27,28,51,53,56,74,75]
APACHE III	4	100%	[11,42,73]	*	[11,42,50,73]	0	0%	*
APACHE IV	3	75%	[47,76]	*	[39,47,76]	1	25%	[49]
SAPS II	6	86%	[7,34,73,75]	*	[7,8,34,62,73,75]	1	14%	[60]
Glasgow Coma Scale (GCS)	5	83%	[47,64]	[5]	[5,39,47,57,64]	1	17%	[49]
Characteristics of admission								
Admission type	10	67%	*	*	[5,7,8,39,42,49,50,57,65,77]	5	33%	[6,34,47,52,64]
Admission Source	5	63%	*	*	[3,5,7,50,64]	3	37%	[6,47,49]
Readmission	4	80%	[3,5,8,50]	*	[3,5,8,50]	1	20%	[47]
LoS before ICU admission	0	0%	*	*	*	4	100%	[3,6,47,49]
ICU type	4	100%	*	*	[6,7,37,61]	0	0%	*
Unable to access GCS	3	100%	[5,47,49]	*	[5,47,49]	0	0%	*
Interventions								
Ventilated	11	92%	[3,5,7,8,42,47,49,57,59,64]	*	[3,5,7,8,39,42,47,49,57,59,64]	1	8%	[6]
Clinical conditions								
Hypomagnesemia	4	50%	[78–81]	*	[78–81]	4	50%	[29,31,82,83]
Delirium	5	71%	[67,84–87]	*	[67,84–87]	2	29%	[55,88]
Malnutrition	7	54%	[89–92]	[93]	[41,63,89–93]	6	46%	[35,77,94–97]
Acute diagnoses								
Gastrointestinal (GI) bleeding	0	0%	*	*		5	100%	[39,47,49,50,56]
Infectious and parasitic diseases	4	100%	[7,8,98]	*	[7,8,39,98]	0	0%	*
APACHE IV diagnoses (Non-operative)								
Trauma	2	67%	*	*	[39,50]	1	33%	[47]
Neurological	1	33%	*	*	[39]	2	67%	[8,47]
APACHE IV diagnoses (Post-operative)								
Trauma	3	100%	[8,47]	*	[8,39,47]	0	0%	*
Chronic Health Items								
Chronic obstructive pulmonary disease (COPD)	3	60%	[3]	*	[3,39,50]	2	40%	[47,49]
Metastatic tumor	2	40%	*	[64]	[50,64]	3	60%	[3,47,49]
Cirrhosis	1	25%	*	*	[39]	3	75%	[47,49,50]
Hepatic failure	2	50%	[49]	[47]	[47,49]	2	50%	[3,50]
Immunosuppression	1	25%	*	[64]	[64]	3	75%	[47,49,50]
Lymphoma	1	25%	*	[64]	[64]	3	75%	[47,49,50]
Acquired immunodeficiency syndrome (AIDS)	0	0%	*	*	*	3	100%	[47,49,50]
Chronic cardiovascular disease	2	67%	[64]	*	[39,64]	1	33%	[47]
Leukemia or myeloma	1	33%	*	[47]	[47]	2	67%	[49,50]
Neoplasms disease	2	67%	*	[37]	[37,39]	1	33%	[49]
Respiratory system disease	2	67%	[37,98]	*	[37,98]	1	33%	[39]
Rhythm disturbance	0	0%	*	*	*	3	100%	[39,47,49]
Reasons for ICU admission								
Sepsis on admission	4	67%	[32,33,99]	*	[32,33,50,99]	2	33%	[3,56]
Heart failure	0	0%	*	*	*	4	100%	[3,47,49,50]
Cardiac arrest	1	33%	[49]	*	[49]	2	67%	[47,50]
Clinical information								
Red blood cell indices (RDW)	4	80%	[51,98,100,101]	*	[51,98,100,101]	1	20%	[102]
Hemoglobin (Hb)	0	0%	*	*	*	3	100%	[51,62,72]
PaO2:FiO2 Ratio	3	100%	[5]	[47,49]	[5,47,49]	0	0%	*
Outcome								
Died in ICU	2	40%	[42]	[64]	[42,64]	3	60%	[8,74,103]

Table 2.1: Summary of factors associated with increased ICU LOS identified by each study.

[Higgins et al., 2003b; Knaus et al., 1993; Nicolas et al., 1987; Straney et al., 2017]. Almost all studies found gender not to be a significant predictor. Over half of the studies that analyzed BMI found that obese patients presented a significantly longer ICU LoS than non-obese ones.

Next, we report on severity scores, such as Acute Physiology and Chronic Health Evaluation (APACHE), Simplified Acute Physiology Score (SAPS), Glasgow Coma Scale (GCS) and Sequential Organ Failure (SOFA). We can note that the great majority of studies found a significant relationship between the severity scores and ICU LoS. Most articles that included APACHE III and IV, SAPS II and III, GCS, and SOFA found them to be significant, and the majority presented a positive effect. APACHE II was the most analyzed severity score, but the results were inconclusive. Two studies compared the influence of both APACHE II and SAPS II in ICU LoS, concluding that the latter had a stronger association [Arabi et al., 2002; Bellia et al., 2019]. Two studies analyzed the influence of both APACHE IV and GCS, noting that the former had a stronger one [Kramer and Zimmerman, 2010; Vasilevskis et al., 2009]. Two studies showed a nonlinear association between the severity scores [Arabi et al., 2002; Knaus et al., 1993] and the ICU LoS: the ICU stay increases as the score increases, but from a certain high score, it starts to decrease.

Regarding the characteristics of ICU admission, most studies found the admission type to be significant: elective surgery patients presented lower LoS compared to medical patients, while emergency surgery patients tend to have a longer ICU stay. The admission source was a significant predictor in most studies: patients from other hospital tend to have a longer ICU LoS, followed by patients from the ward, emergency room and operating recovery room. The majority of studies found that readmitted patients tend to have a longer ICU stay. In all four studies, hospital LoS before ICU admission was not a significant factor. The ICU type was always found to be significant, but the results were inconclusive regarding which type provided longer stay. In all three studies, patients unable to access GCS due to sedation or paralysis presented statistically greater ICU stay.

Some studies analyzed factors related to clinical conditions. The influence of hypomagnesemia, coma, and malnutrition was uncertain. Regarding the influence of delirium, most studies found it to be positively related to ICU LoS. The two studies of hypernatremia found that patients with this condition tend to have a longer ICU stay. Twelve articles analyzed the effect of interventions, such as mechanical ventilation, and the vast majority found that ventilated patients presented higher ICU stay.

Regarding the acute diagnoses, all gastrointestinal (GI) bleeding studies

did not find it to be a significant predictor. Patients with infectious diseases and with cerebrovascular accident were found to have a greater ICU stay. Some studies also analyzed the APACHE IV diagnoses for both non-operative and post-operative patients. Most studies found trauma and respiratory diagnoses to be positively related to ICU LoS for both types of patient.

Next, we report on factors related to chronic health items. The majority of studies found the following diseases to be positively associated with ICU LoS: chronic obstructive pulmonary disease (COPD), chronic cardiovascular disease, and respiratory system disease. Regarding the reasons for ICU admission, sepsis was the most analyzed one, and most studies found that sepsis patients tend to have greater ICU stay. Moreover, all studies found myocardial infarction, intracerebral hemorrhage, pulmonary edema, and subarachnoid hemorrhage to be positively related to ICU LoS.

Some articles analyzed the influence of clinical information in the ICU LoS. The majority of studies found that red blood cell indices, albumin-creatinine ratio, body temperature and MR-proANP were positively associated with ICU stay. PaO2:FiO2 ratio was a significant factor, but negatively related to ICU LoS. The influence of mortality and organizational factors in the ICU LoS were inconclusive.

2.3.5 Meta-analysis

We included 28 studies reporting appropriate statistics that could be converted to one of the following effect sizes: partial correlation (rp) and standardized mean difference (d). Six factors reported appropriate statistics and were meta-analyzed. The effect size for age, gender and mechanical ventilation was the partial correlation (Figure 2.2), and for hypomagnesemia, delirium, and malnutrition we used the standardized mean difference (Figure 2.3). The overall quality of the included studies was high: 20 articles (71%) met at least five of the six QUIPS domains analyzed, and 27 articles (96%) attended at least four domains.

Five studies reported statistics for age and the meta-analysis demonstrated non-significant association with ICU LoS (Rp= 0.02; 95%CI: -0.03, 0.07; p-value = 0.44; I2 = 95.46%) [Choi and Lee, 2016; Ely et al., 2001; Moran et al., 2008; Reini et al., 2012; Vasilevskis et al., 2009]. The meta-analysis for gender (reference = female) also demonstrated non-significant relationship with ICU LoS (Rp= 0.005; 95%CI: 0.00, 0.01; p-value = 0.05; I2 = 2.61%) [37,42,52]. Regarding the meta-analysis for mechanical ventilation (reference = no), we noted a significant and positive association with ICU LoS (Rp=

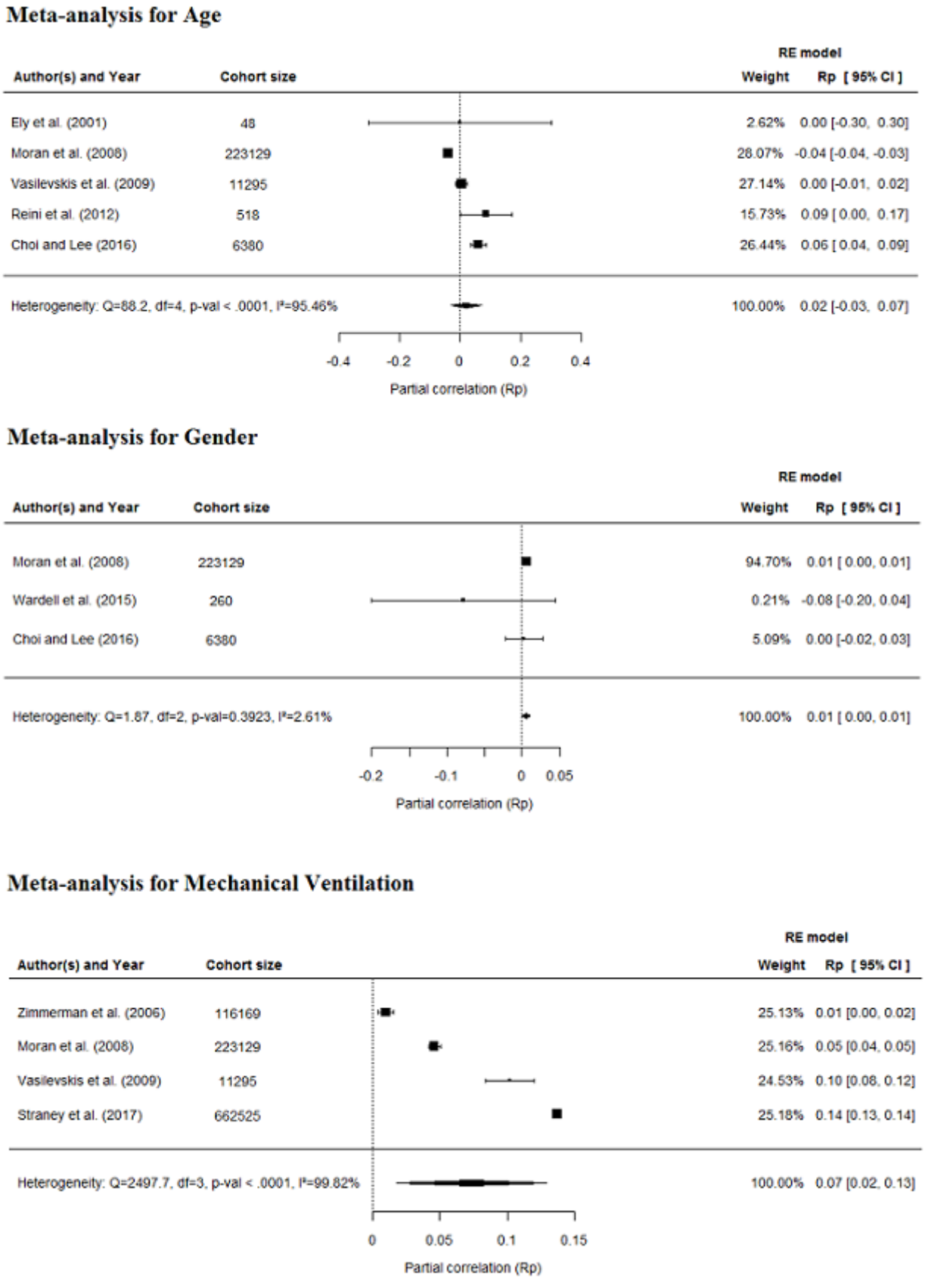


Figure 2.2: Meta-analysis for the impact of age, gender and mechanical ventilation in ICU length of stay.

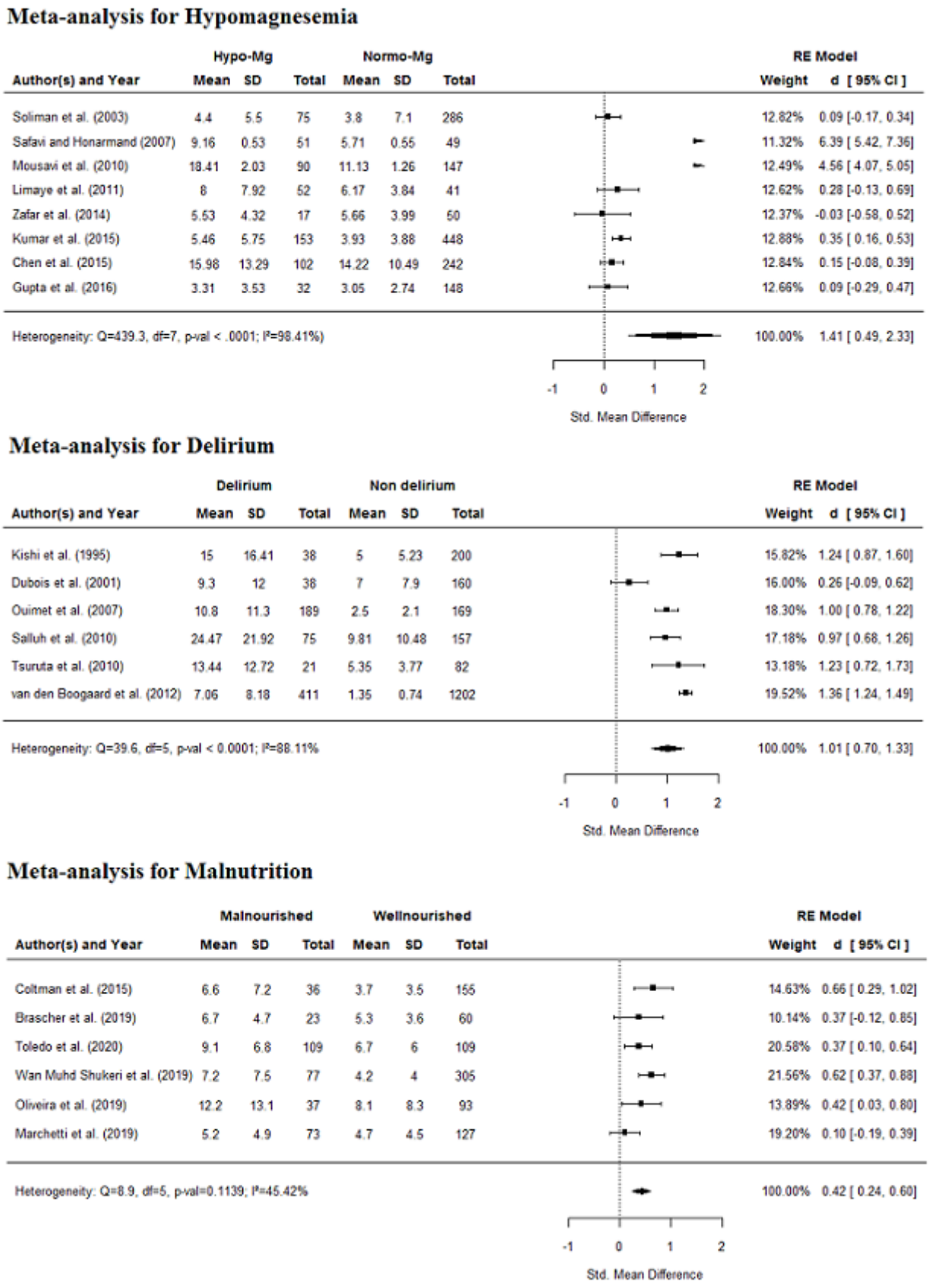


Figure 2.3: Meta-analysis for the effect of hypomagnesemia, delirium and malnutrition in ICU length of stay.

0.07; 95%CI: 0.02, 0.13; p-value = 0.0097; I2 = 99.82%) [Moran et al., 2008; Straney et al., 2017; Vasilevskis et al., 2009; Zimmerman et al., 2006].

The meta-analysis for hypomagnesemia showed a significant and positive association with ICU LoS (d = 1.41; 95%CI: 0.49, 2.33; p-value = 0.0026; I2 = 98.41%): hypomagnesemia patients tend to stay 1.41 days longer. The meta-analysis for delirium showed also a significant and positive relationship with ICU LoS (d = 1.01; 95%CI: 0.70, 1.33; p-value < 0.0001; I2 = 88.11%) [67,84–88]: delirium patients tend to stay 1.01 days longer. The meta-analysis for malnutrition showed a significant but weaker relationship with ICU LoS (d = 0.42; 95%CI: 0.24, 0.60; p-value < 0.0001; I2 = 45.42%) [35,90–92,97,104]: malnourished patients (NUTRIC score >= 5) tend to stay 0.42 days longer.

Publication bias was detected by Begg's test [Begg and Mazumdar, 1994]. There was no evidence of publication bias for all meta-analyses considering 5% of significance level.

2.3.6 Sensitivity analysis

To assess the robustness of the estimates' stability, a sensitivity analysis was performed after excluding one study at a time. We found heterogeneous studies for the meta-analysis of age, hypomagnesemia, and delirium. After excluding those studies, the heterogeneity was significantly reduced (Figure 2.4). None of the results in terms of statistical significance was altered. Regarding the mean estimate, we noted a significant change only for hypomagnesemia.

The meta-analysis for age remained to show a non-significant association with ICU LoS (Rp= 0.04; 95%CI: 0.00, 0.09; p-value = 0.08; I2 = 78.3%) [Choi and Lee, 2016; Ely et al., 2001; Reini et al., 2012; Vasilevskis et al., 2009]. For hypomagnesemia, the results remained significantly associated with ICU LoS. The heterogeneity decreased from 98.41% to 0%, and the mean effect size reduced from 1.41 to 0.21 (d = 0.21; 95%CI: 0.10, 0.32; p-value = 0.0003; I2 = 0%) [Chen et al., 2015; Gupta et al., 2016; Kumar et al., 2015; Limaye et al., 2011; Soliman et al., 2003; Zafar et al., 2014]. Regarding the sensitivity analysis of delirium, the results also remained significantly associated with ICU LoS, and the heterogeneity decreased from 88.11% to 0% (d = 1.05; 95%CI: 0.90, 1.20; p-value < 0.0001; I2 = 0%) [Kishi et al., 1995; Ouimet et al., 2007; Salluh et al., 2010; Tsuruta et al., 2010].

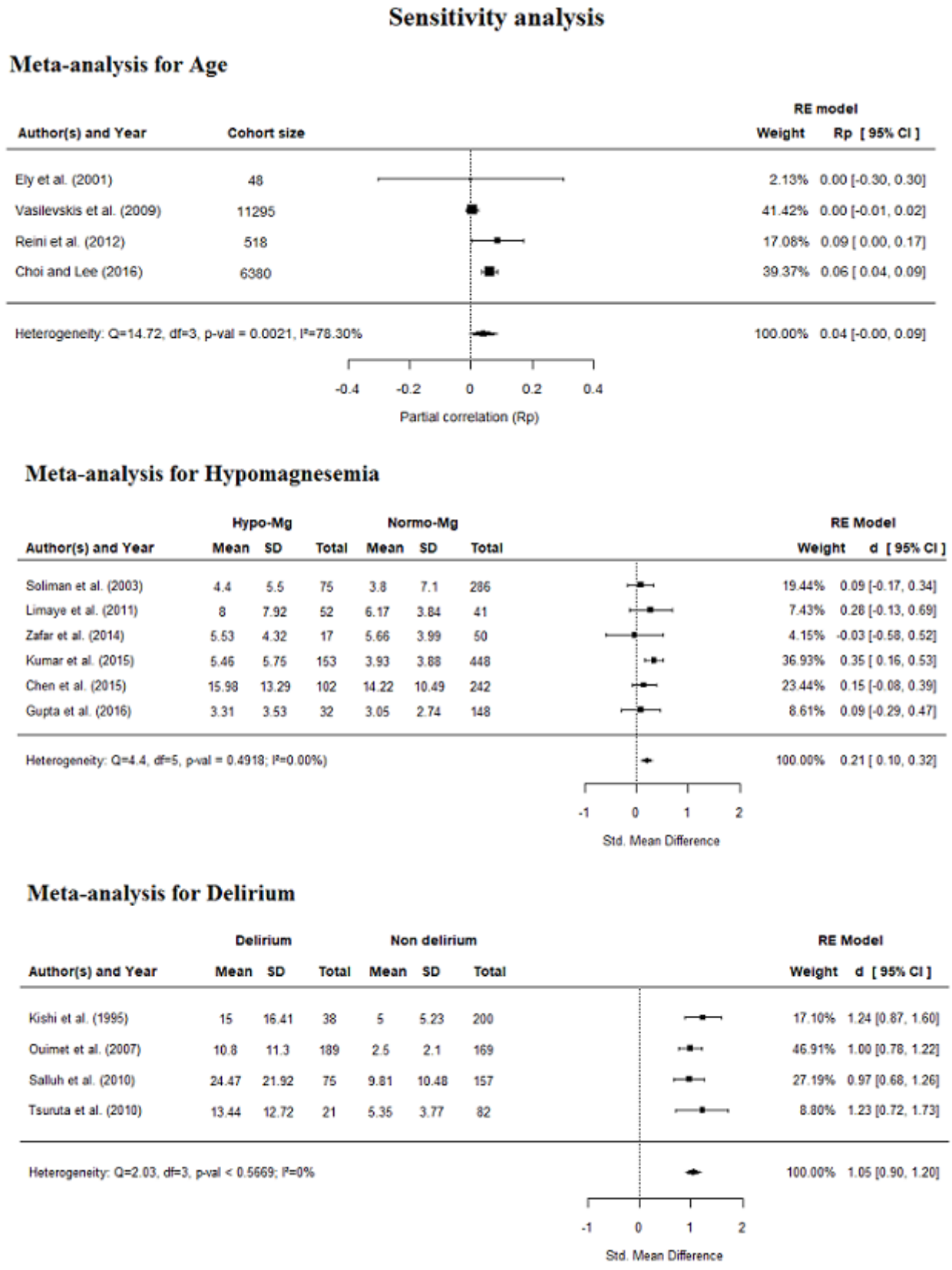


Figure 2.4: Sensitivity analysis of the meta-analysis.

2.4 Discussion

This work systematically reviewed 113 papers dealing with risk factors of ICU LoS. We performed an extensive analysis of risk factors that should be included in future prediction models to improve their predictive capacity. We also performed a meta-analysis of six factors from 28 articles. The meta-analyses concluded that patients with mechanical ventilation, hypomagnesemia, delirium, and malnutrition tend to have higher ICU stay, and showed that age and gender were not significantly associated with an extended stay. Table 2.2 summarizes our findings regarding risk factors more frequently associated with a higher ICU stay.

List of risk factors of higher ICU stay
Patient demographics, severity scores and characteristics of admission:
Patients with higher Body Mass Index (obese patients); with higher severity scores (APACHE, SAPS, or GCS); admitted for emergency surgery; coming from another hospital; readmitted; unable to access GCS score; and with mechanical ventilation.
Clinical conditions and diagnoses:
Hypomagnesemia, delirium, malnutrition, infectious diseases, cerebrovascular accident, trauma, and respiratory diagnoses.
Chronic Health Items:
Chronic obstructive pulmonary disease (COPD) and other respiratory system disease; and chronic cardiovascular disease.
Reasons for ICU admission:
Sepsis, intracerebral hemorrhage, myocardial infarction, pulmonary edema and subarachnoid hemorrhage.
Clinical information:
High levels of red blood cell, body temperature, albumin-creatinine ratio, and MR-proANP; and lower PaO2:FiO2 ratio.

Table 2.2: Summary of potential risk factors of ICU stay.

This study analyzed 89 possible risk factors of ICU LoS and reported a summary list of relevant ones, as presented in Table 2.2. We recommend future studies related to the prediction of ICU stay to include at least these risk factors in their models. Accurate prediction models can bring the information whether a patient may have a short or prolonged ICU LoS, which can help the decision-makers to act accordingly. Patients with a very short ICU stay means that the unit has the potential to release the bed in a few days. This information is important to the managers, especially if the ICU is crowded. On the other hand, a long ICU LoS means that the patient may hold the bed for a prolonged period. So, this information can be used by managers to change the protocol of care for this patient, trying to reduce the time in the

ICU. Next, we discuss our findings in light of previous reviews. Of note, the existing reviews deal with a specific population [Atashi et al., 2018; Seaton et al., 2016] or focus on models used to predict LoS [Atashi et al., 2018; Awad et al., 2017; Verburg et al., 2017] and, to the best of our knowledge, no updated SLR and meta-analysis exist for factors associated with ICU LoS considering general population. Therefore, we compared our results with a range of specific reviews.

Jiang et al. [2017] performed a meta-analysis for the impact of hypomagnesemia on ICU outcomes and concluded that ICU stay for hypomagnesemia group was 1.85 days longer. We updated the previous meta-analysis, and our results were in line with theirs. Zhang et al. [2013] performed a meta-analysis to evaluate the impact of delirium in critically ill patients. The meta-analysis included ten studies, some of them from specific ICU population, and the results showed that delirious patients tend to stay 7.32 days longer. We performed a meta-analysis considering only studies from general ICU population, and our results also demonstrated a significant association. However, the mean effect was shorter ($d = 1.01$ days).

Zhang et al. [2015] performed a meta-analysis to assess the effects of gastric tonometry guided therapy on patient outcome in ICUs and suggested that it could not significantly reduce the days spent in the ICU. Muscedere et al. [2017] performed a meta-analysis regarding the impact of frailty on ICU outcomes and demonstrated a non-significant relationship with ICU LoS. Chant et al. [2011] evaluated the effect of catheter-associated urinary tract infection (CAUTI) in critically ill patients through a meta-analysis and demonstrated a significant increase in the ICU stay of 2.4 days in CAUTI patients. We did not find any other new study to update those three previous meta-analyses.

Regarding the severity scores, almost all studies that analyzed APACHE IV and SAPS III found them to be significant predictors of ICU stay. Since both severity scores consider in their formulations other potential predictors, like the ICU admission reason, it may explain the significance of those scores. The previous reviews on this topic suggested that APACHE IV provides better predictions compared to SAPS III. APACHE IV includes 116 specific ICU admission reasons, whereas SAPS III includes only ten. Since the admission reason is one of the main predictors of outcome, it may explain the accuracy of APACHE IV model [Breslow and Badawi, 2012b; Keegan et al., 2011]. We did not find case studies comparing those two scores, and we suggest future studies to include those scores and present their comparison results.

We analyzed two studies that showed a nonlinear association between

the severity scores [Arabi et al., 2002; Knaus et al., 1993] and the ICU stay: the LoS increases as the score increases, but from a certain score it starts to decrease. Two reviews of severity scores presented the same behavior and noted that it could be explained by the greater mortality rate in the highest severity levels [Breslow and Badawi, 2012a; Higgins, 2007].

Four articles analyzed large datasets and reported their results regarding the relative importance of each variable for ICU LoS prediction [Knaus et al., 1993; Kramer and Zimmerman, 2010; Weissman et al., 2018; Zimmerman et al., 2006]. The following factors were commonly reported as relevant predictors: APACHE scores, GCS, PaO2:FiO2 ratio, mechanical ventilation, reason for ICU admission, inability to access GCS, and ICU admission source. Of note, those results are in line with our findings.

This work has the following limitations. Regarding the systematic review, we did not include non-English written articles in our literature search because it would complicate the process of title, abstract, and full-text screening, requiring a translation pre-process, and could prevent the process of consensus-building among authors when disagreements on exclusion of studies arises. Regarding the quality of studies, most of them considered a cohort with both medical and surgical patients. However, there was a range of inclusion and exclusion criteria, which could generate cohorts with more severe or less severe patients. Furthermore, only 38 studies adjusted their models for potential confounders, which may also affect our results. Regarding the meta-analysis, the high heterogeneity found for some studies could be partly explained by the differences in clinical characteristics of individual studies (settings) and study designs (retrospective vs. prospective). Therefore, we performed a random-effects meta-analysis by assuming that the true effects were normally distributed. In this model, more weight is assigned to small-sized studies compared to the fixed-effects model [Thompson and Sharp, 1999]. Moreover, we did a sensitivity analysis to assess the robustness of the estimates' stability. After excluding the high heterogeneous studies, we concluded that none of the results in terms of statistical significance was altered, and we noted a significant change only for the estimate of hypomagnesemia.

To make possible more reviews and meta-analyzes in this topic, we recommend future studies to include the statistics for all variables analyzed, and not only for significant ones. Moreover, studies should report not only p-values but also the effect-sizes (i.e., regression coefficients, confidence interval and p-values). From 113 studies, we could only meta-analyze 28 of them, because the majority did not present appropriate statistics, which is also a limitation of this work. So, only six factors were meta-analyzed from 89 that

could have been if we had the appropriate statistics. It would be of great value to perform a meta-regression related to predictors of ICU LoS, but we need future studies to report the complete statistics.

2.5 Conclusions

This work systematically reviewed and meta-analyzed papers dealing with risk factors of ICU stay and suggested a list of factors that should be considered in prediction models for ICU LoS. In summary, the main risk factors that should be considered in future prediction models are, as follows: severity scores, BMI, admission source, admission type, readmission (yes/no), inability to access GCS (yes/no), mechanical ventilation (yes/no), clinical conditions (hypomagnesemia, delirium, malnutrition, infectious diseases, cerebrovascular accident, trauma, and respiratory diagnoses), chronic health items (COPD and chronic cardiovascular disease), reasons for ICU admission (sepsis, intracerebral hemorrhage, myocardial infarction, pulmonary edema and subarachnoid hemorrhage) and clinical information (levels of red blood cell, body temperature, MR-proANP, albumin-creatinine ratio, and PaO2:FiO2 ratio). Our findings can be used by future prediction models to improve their predictive capacity of prolonged stay patients, which can assist in planning the number of resources required, driving quality improvement actions, and enabling case-mix adjustments for benchmarking analysis.

3
Data-driven methodology to predict ICU length of stay

This chapter presents a structured data-driven methodology to approach the main demands for ICU managers regarding the prediction of ICU LoS: (i) planning the number of beds and staff required to fulfill the need for ICU care (Section 3.2); (ii) identifying patients with a high risk of prolonged ICU LoS to drive immediate quality improvement (Section 3.3); and (iii) enabling case-mix correction when comparing the LoS between ICUs (benchmarking) (Section 3.4). Figure 3.1 shows the proposed framework, which summarizes the topics covered in this chapter.

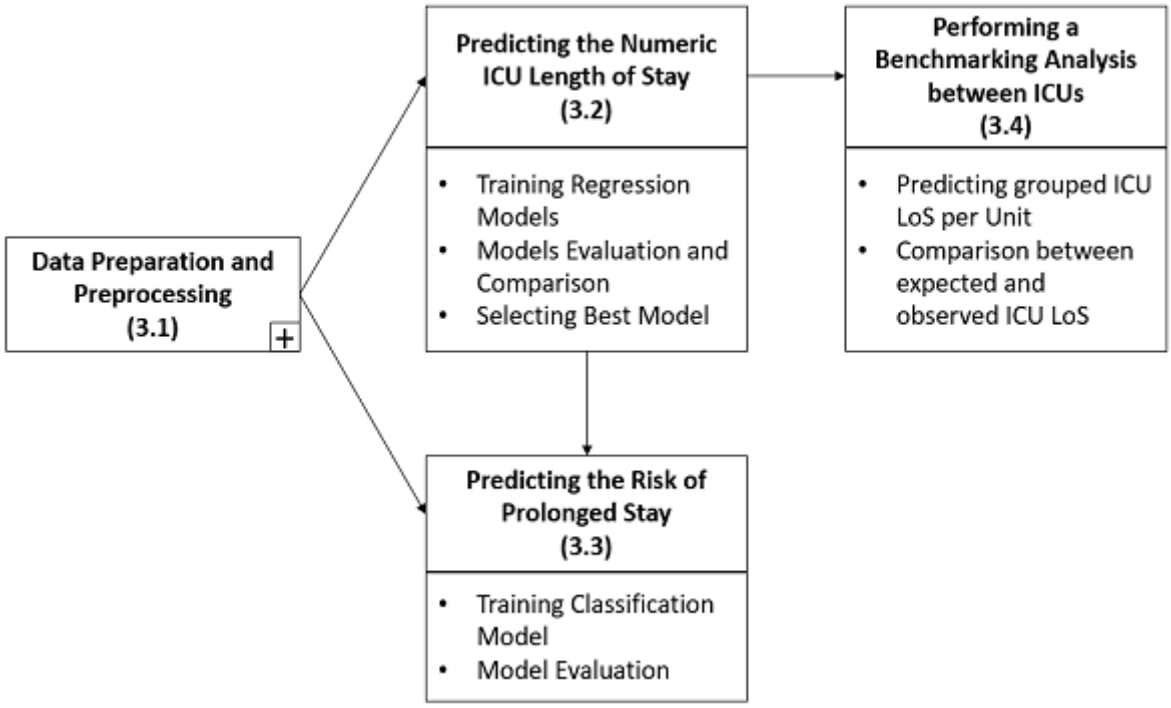


Figure 3.1: Framework of ICU LoS prediction topics covered in this chapter

First, we will preprocess the dataset, applying transformations that can improve future models performance (Section 3.1). Second, after preprocessing the dataset, we can apply the methodology to predict the numeric ICU length of stay (Section 3.2), which is the leading information necessary for planning the ICU resources. This section will show the topics of training and evaluating regression models. Third, we will use the best type of model found applying

the methodology of Section 3.2 to train a classification model to predict the patient's risk of being a prolonged stay (Section 3.3), which helps to identify possible prolonged stay patients. This section presents the particularities that should be considered when using classification models to ICU LoS. Finally, we will use our best model of Section 3.2 to predict the grouped length of stay for each ICU and then present a methodology to perform a non-biased benchmarking analysis between ICUs (Section 3.4).

3.1 Data Preparation and Preprocessing

Figure 3.2 presents the framework of data preparation and preprocessing methodology. Each step is explained in detail considering the prediction of the numeric ICU length of stay. The framework includes the following steps:

- Data Preparation: Import the database, understand the characteristics, and propose some new features based on the existing ones (feature engineering);
- Visualization and Data Cleaning: Analysis of missing values, descriptive analysis, and outliers detection and treatment;
- Data Splitting: Splitting the data into training and testing;
- Data Preprocessing: Treating the dataset before the application of regression models.

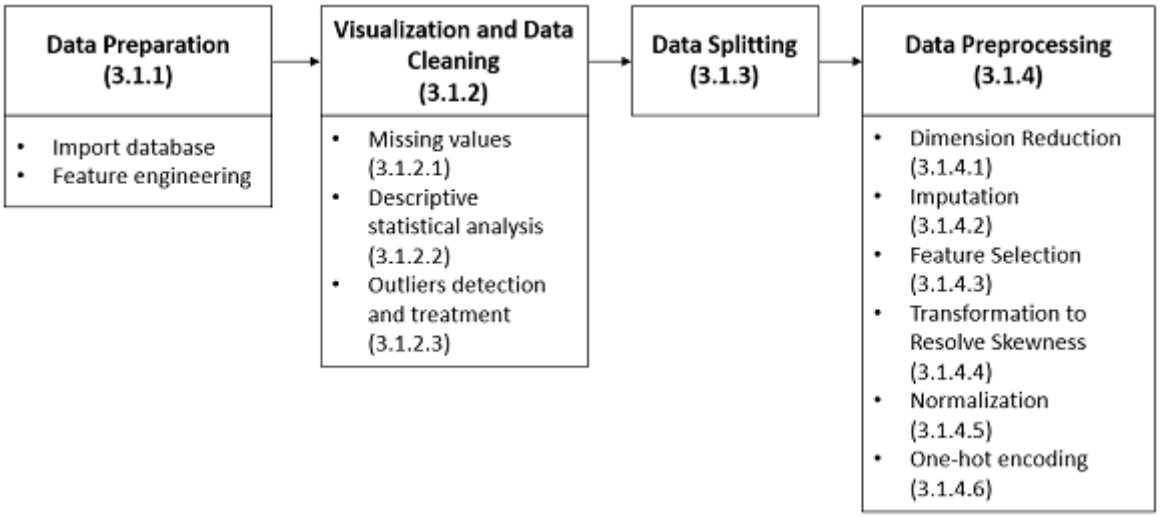


Figure 3.2: Framework of the Data Preparation and Preprocessing methodology

3.1.1 Data Preparation

In this section, we will present the data preparation. First, we will import the database and observe the features. After that, we will analyze which feature can be created based on the existing ones. This is an important step called "feature engineering".

Some authors proposed general definitions for feature engineering. Kuhn and Johnson [2019] defines it as a process that includes the following topics: the transformation of a predictor, interaction of two or more predictors such as a product or ratio, the functional relationship among predictors, equivalent re-representation of a predictor. Feature transformation is the process of constructing new features from existing ones [Dong and Liu, 2018]. According to Morid et al. [2019], using temporal variables with greater granularity (e.g., each invasive procedure taken in the past six months) could improve the prediction model's performance since these variables expose more information than using aggregated variables (e.g., the sum of procedures in the past six months). Therefore, in this step of the methodology, we will propose new features from existing ones.

3.1.2 Visualization and Data Cleaning

This section will present the strategy to: treat missing values, show the descriptive statistical analysis, and detect and treat outliers.

3.1.2.1 Missing values

Our study follows the guidelines for reporting analysis potentially affected by missing data [Chevret et al., 2015; Sterne et al., 2009], which recommend reporting the missing data structure, the multiple imputation approach, and comparing the imputed and the observed data (if the variable has a large proportion of missing). Features with more than 30% of missing must be excluded from the analysis. We will present the complete treatment for missing data in the "Imputation" section.

3.1.2.2 Descriptive statistical analysis

After the data preparation, we recommend presenting a descriptive analysis of the data. This analysis aims to understand the variables presented

in the dataset and provide insights into factors related to prolonged length of stay.

3.1.2.3 Outlier detection and treatment

A common approach to visualize univariate extreme values is the use of box plots. In a boxplot, the statistics of a univariate distribution are summarized in terms of five quantities: "minimum/maximum" (whiskers), the upper and lower quartiles (boxes), and the median (line in the middle of the box) [Aggarwal, 2017]. The distance between the upper and lower quartiles is referred to as the interquartile range (IQR). If there are points bigger than 1.5 IQR (above the third quartile and bellow the first quartile), they are considered moderate outliers. If there are points higher than 3 IQR, they are considered extreme outliers.

There are basically three approaches for treating outliers in a dataset [Kwak and Kim, 2017]: (i) removing outliers by trimming the dataset; (ii) replacing the values of outliers with expected values or reducing the influence of outliers through outlier weight adjustments; (iii) estimate the values of outliers using robust techniques. The first and second approaches may not be statistically valid in general, and they can lead to serious bias. These biases can be overcome considering the third approach and using multiple imputation methods to estimate the value of outliers [Sterne et al., 2009]. Therefore, for the numeric covariates of the model, we recommend analyzing the presence of extreme outliers with boxplots and then applying the third treatment approach. A detailed explanation about multiple imputation methods will be presented in Section 3.1.4.2 (Imputation section).

Regarding the dependent variable (ICU length of stay), because of the non-normality of its distribution, truncation of the data or deleting outliers have been undertaken before possible data transformation. Studies implementing the APACHE III and IV algorithms for ICU LoS prediction used the truncation at 30 days (99% percentile) to treat the outliers, truncating 1% of the data [Kramer and Zimmerman, 2010; Niskanen et al., 2009; Vasilevskis et al., 2009; Verburg et al., 2014; Zimmerman et al., 2006]. Therefore, for the dependent variable, we recommend following the literature not removing the outliers and applying the truncation at high percentiles (e.g., 99% or 95%).

3.1.3 Data Splitting

Before data preprocessing and modeling, we have to decide which samples will be used to evaluate performance. According to Kuhn and Johnson [2013], to provide an unbiased sense of model effectiveness, we should evaluate the predictive model on samples that were not used to build or tune the model. The "training" dataset is the general term for the samples used to create the model, while the "testing" dataset is used to measure performance.

In order to split the dataset in training and test, it is recommended to use stratified random sampling, which applies random sampling within subgroups (such as the classes) [Kim, 2009; Molinaro et al., 2005]. This strategy accounts for the outcome when splitting the data, providing a higher likelihood that the outcome distributions will match. When the outcome is a number (ICU LoS), we can use a similar approach; the numeric values are broken into groups (e.g., low, medium, and high), and the randomization is executed within these groups. Therefore, we recommend applying this sampling methodology, splitting 80% of the dataset for training and 20% for testing (as proposed by Kuhn and Johnson [2013]).

3.1.4 Data Preprocessing

Data preprocessing techniques refer to transformations of the training and testing datasets to improve model performance. Transformations can be used to reduce the impact of data skewness or outliers, to remove predictors based on their lack of information content, or to extract new features. The need for data preprocessing is determined by the type of model being used. Kuhn and Johnson [2013] proposed the main preprocessing steps that should be considered before applying prediction models, which can be summarized by the following steps: dimension reduction, imputation, transformations to resolve skewness, normalization, and one-hot encoding. Next, we will present each one in detail.

3.1.4.1 Dimension reduction

There are potential advantages to removing predictors before modeling. First, fewer predictors mean lower computational time and complexity. Moreover, removing zero variance (and near zero) predictors and correlated predictors can improve the model's performance. These transformations might lead to a more parsimonious and interpretable model [Kuhn and Johnson, 2013].

- Zero and Near-Zero Variance Predictors

A predictor that has a single unique value is considered a zero variance predictor. This variable does not add any additional information to the models and should be removed. Some models are invariant to this variable (like tree-based models); however, other models can have problems in the computations (like linear regression). Another type of problematic predictor is the near-zero variance predictor, which has the vast majority of cases presenting a unique value and few cases showing other values. A rule of thumb for detecting near-zero variance predictors is: the fraction of unique values over the sample size is low ($\leq$10%); and the ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value is large ($\geq$20). If both criteria are true, these variables should be removed from the model [Kuhn and Johnson, 2013]. We recommend excluding zero and near-zero variance features.

- Identifying Correlated Predictors

The term "collinearity" refers to the situation where a pair of predictors have a substantial correlation with each other. There are good reasons to avoid data with high collinearity. First, redundant predictors can add more complexity to the model than information. Second, we commonly have a cost and time associated when obtaining each predictor data; therefore, fewer variables are better for the model. Moreover, there are mathematical disadvantages to having correlated predictors. For instance, using highly correlated predictors in techniques like linear regression can result in highly unstable models, numerical errors, and degraded predictive performance. A statistic called the variance inflation factor (VIF) can be used to identify predictors that are impacted [Myers and Myers, 1990]. However, this method may be inadequate for several reasons: it was developed for linear models, it requires more samples than predictor variables, and, while it does identify collinear predictors, it does not determine which should be removed to solve the problem [Kuhn and Johnson, 2013]. A recommended approach to dealing with this issue is to remove the minimum number of predictors to ensure that all pairwise correlations are below a certain threshold. While this method only identifies collinearities in two dimensions, it can positively affect the performance of some models [Kuhn and Johnson, 2013]. Therefore, we recommend considering this approach, which has some differences according to the type of feature. For numeric variables, we recommend using Pearson Correlation with a threshold of 0.75 [Kuhn and Johnson, 2013]. For categorical ones, we suggest employing Cramér's V with a threshold of 0.5 [Cohen, 1988].

3.1.4.2 Imputation

Missing data occur in almost all medical research. Inadequate handling of them can lead to biased or inefficient estimates of parameters and incorrect confidence intervals. In all statistical analyses, some assumptions are made about the missing data [White et al., 2011]. Little and Rubin [2019] proposed a framework to classify the missing data, which can be (i) missing completely at random (MCAR — the probability of data being missing does not depend on the observed or unobserved data), (ii) missing at random (MAR — the probability of data being missing does not depend on the unobserved data, conditional on the observed data) or (iii) missing not at random (MNAR — the probability of data being missing does depend on the unobserved data, conditional on the observed data). The distinction between MAR and MNAR can not be made from the dataset alone. However, the MAR assumption can be more plausible by collecting more explanatory variables and including them in the analysis [White et al., 2011].

When it is plausible that data are missing at random (MAR) but not completely at random (MCAR), analyses based on complete cases (excluding missing values) may be biased. A variety of ad hoc approaches are commonly used to deal with missing data: replacing missing values with the mean of the observed values, with the last measured value, or using a missing category indicator. None of these approaches is statistically valid in general, and they can lead to serious bias. These biases can be overcome using multiple imputation methods, which allow individuals with incomplete data to be included in analyses [Sterne et al., 2009].

In large data sets, it is common to occur missing values in several variables. Multiple imputation by chained equations (MICE) is a practical approach to generating imputations based on a set of imputation models, one for each variable with missing values. An advantage of MICE is the ability to handle different variable types (continuous, binary, unordered categorical, and ordered categorical), because each variable is imputed using its own imputation model [White et al., 2011]. The MICE algorithm works as follows. First, all missing values are filled in by simple random sampling with replacement from the observed values. The first variable with missing values (x1) is regressed on all other variables (x2, . . . , xk), restricted to individuals with the observed x1. Missing values in x1 are replaced by simulated values from the posterior predictive distribution of x1. Then, the next variable with missing (x2) is regressed on all other variables (x1, x3, . . . , xk), restricted to individuals with the observed x2, and using the imputed values of x1. As done before,

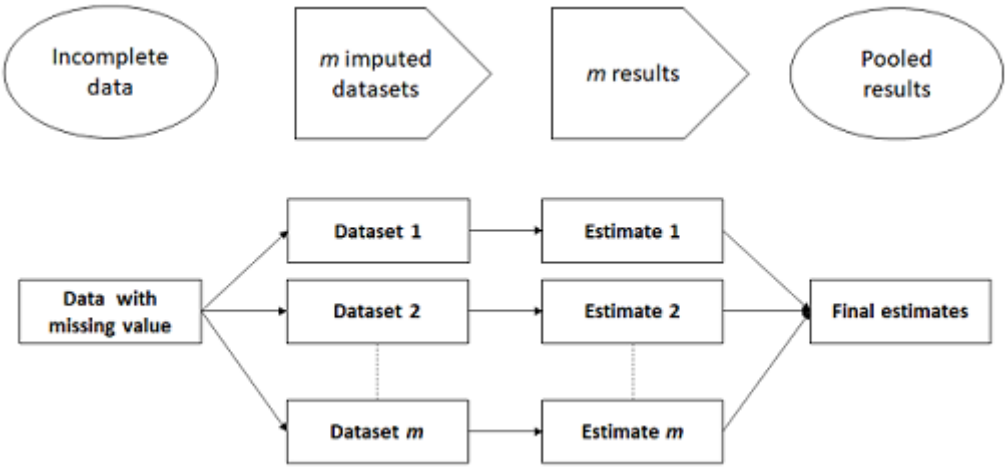


Figure 3.3: Multiple imputation steps

the missing values in x2 are replaced by draws from the posterior predictive distribution of x2. This process is repeated for all variables with missing values. In order to stabilize the results, the whole process is repeated m times to give m imputed data sets [Chevret et al., 2015; White et al., 2011]. Rubin's rules give overall estimates and corresponding standard errors from the m separate analyses [Rubin, 2004]. The whole procedure is resumed in Figure 3.3.

Standard texts on multiple imputation suggest that small numbers of imputed data sets (m = 3 or 5) are adequate. Graham et al. [2007] argued that we should instead choose the number of imputations to limit the loss in power for testing an association of interest. To limit the loss in power to no more than 1%, they recommended m greater than 20. More recent advice is that m should be at least equal to the percentage of incomplete cases [Chevret et al., 2015; White et al., 2011]. Moreover, White et al. [2011] do not recommend imputation in variables with more than 30% of missing, because it can add imperfections in the imputation procedure.

Therefore, we suggest applying the MICE imputation algorithm to the variables with incomplete data considering m at least equal to the percentage of incomplete cases. Variables with fraction of incomplete data greater than 30% should not be included in the study.

3.1.4.3 Feature Selection

The feature selection has the objective to reduce the dimension of the problem by removing variables that do not present a significant contribution to the model. We recommend testing the Recursive Feature Elimination (RFE) with random forest (RF-RFE), and the RFE with Bagging (Treebag-RFE), as

proposed by Kuhn and Johnson [2013].

3.1.4.4 Transformation to Resolve Skewness

Several literature studies noted skewness in numerical ICU features from different datasets and had to use transformations to treat this problem [Caetano et al., 2014; Choi and Lee, 2016; Li et al., 2019; Moran et al., 2008; Moran and Solomon, 2012; Niskanen et al., 2009; Straney et al., 2017; Verburg et al., 2014, 2018c; Zimmerman et al., 2006]. Those studies also showed that the outcome variable (ICU LoS) demonstrated a markedly right skewed distribution and most of them used log-transformations for this problem.

A general rule of thumb to consider is that skewed data whose ratio of the highest value to the lowest value is greater than 20 have significant skewness. Therefore, replacing that data with the log, square root, or inverse may help to remove the skew [Kuhn and Johnson, 2013]. However, it is important to find which one of those statistical methods would be better to get the appropriate transformation. Box and Cox [1964] propose a family of transformations that are indexed by a parameter, denoted as λ:

$$x^* = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x) & \text{if } \lambda = 0 \end{cases} \tag{3-1}$$

In addition to the log transformation, this family can identify square transformation (λ = 2), square root (λ = 0.5), inverse (λ = -1), and others in-between. Box and Cox [1964] used maximum likelihood estimation to determine λ in training data. We suggest applying this procedure independently to each numeric predictor that has significant skewness.

3.1.4.5 Normalization

The data normalization refers to set the numerical variables of the database on a common scale. These transformations are generally used to improve the numerical stability of some calculations. Normalization methods affect differently on different classifiers. Distance-based classifiers like SVM, KNN, and neural networks dramatically benefit from normalization [Kuhn and Johnson, 2013]. Other length of stay prediction studies also used the normalization in their data [Caetano et al., 2014; Li et al., 2019; Liu et al., 2006]. A common method to normalize the dataset is to scale the data between 0 and 1, also known as "Min-Max Normalization" or "Normalization by Range". This method is demonstrated in Equation 3-2, where $x = (x_1, ..., x_n)$ and z_i

is the ith normalized data. We recommend using this method to scale the numerical variables.

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{3-2}$$

3.1.4.6 One-hot encoding

One-hot encoding is a method used to handle datasets with mixed data types (numerical, categorical, and binary) since some machine learning prediction models do not accept this type of data. The method is used to encode a categorical feature with k possible values to k features. The feature representing the corresponding category has a value of 1, and all other features have values of 0.

3.2 Predicting the Numeric ICU Length of Stay

There are a set of prediction studies to the length of stay in the Intensive Care Unit [Caetano et al., 2014; Li et al., 2019; Liu et al., 2006; Weissman et al., 2018]. A range of models was tested, but none presented a structured methodology to develop prediction models. This section aims to develop a structured data-driven methodology for numeric ICU LoS prediction. The objective is to predict the patient length of stay in a specific ICU admission based on the first 24h data, which helps planning the number of resources required to fulfill the need for ICU care. Despite not being our aim, it is important to note that another possible objective would be to predict the total patient time in all ICUs, which considers the multiple patient admissions. In this thesis, we will consider the ICU LoS prediction for each independent ICU admission.

A recent literature review analyzed six prediction articles that applied and compared different regression models to predict ICU LoS [Peres et al., 2021]. The authors noted that Support Vector Regression (SVR), Gradient Boosting Machine (GBM), and Random Forests (RF) presented superior results compared to other data-driven models. Therefore, we recommend testing the following regression models, as implemented in the caret package [Kuhn, 2009]: Nonlinear Regression models, like SVR and k-Nearest Neighborhood (kNN); Tree-based models, such as GBM, CART, RF, and Bagging; and Linear Regression models, like Linear Regression (LR), and Generalized Linear Model (GLM) with Negative Binomial distribution.

Support Vector Machines (SVMs) [Cortes and Vapnik, 1995] are sparse

kernel machines, a type of models that rely only on a subset of data, the support vectors. SVMs allows the use of kernels to project the input data to a higher-dimensional space. The model separates the training data by means of a good-fitting hyperplane into two classes. Kernels can be used to transform this hyperplane into a nonlinear input separator, making it a very effective classifier. Support vector regression (SVR) is the application of SVMs to regression, in which a linear function is fit through the training set. Also, kernels can be used to transform the linear fit to a nonlinear curve.

K-Nearest Neighbors [Altman, 1992] is a learning algorithm capable of regression as well as classification. In the kNN model, a new sample is predicted based on the training set's k-closest data points. The question remains as to how many neighbors should be used since too few neighbors may generate over-fitting while too many may not be sensitive enough to achieve reasonable performance. In the kNN regression, the model considers the average neighbor value, while in classification the mode of the class of the k-nearest neighbors is used.

Tree-based models are learning algorithms that performs one or more if-then statements for the predictors to partition the data. These models generalize training data by building a tree structure. Within the tree partitions, a model is used to predict the outcome [Kuhn and Johnson, 2013]. There are many techniques for constructing regression trees. One of the oldest and most utilized is the Classification and Regression Trees (CART) [Breiman et al., 1984]. CART is a decision tree algorithm capable of regression as well as classification. For regression, the model begins with the entire data set, S, and searches every distinct value of every predictor to find the predictor and split value that partitions the data into two groups (S1 and S2) such that minimizes the overall prediction error measure. Then, within S1 and S2, this method searches for the predictor and split value that best reduces error. So, because of the recursive splitting nature of regression trees, this method is also known as recursive partitioning.

Bootstrap Aggregation (Bagging) was one of the earliest developed ensemble techniques [Breiman, 1996]. Bagging is a general approach that uses bootstrapping in conjunction with any regression (or classification) model to construct an ensemble. Each model in the ensemble is then used to generate a prediction for a new sample, and these m predictions are averaged to give the bagged model's prediction. Bagged models provide advantages over models that are not bagged. While the "simple" CART decision tree used to produce unstable predictions, using bagged trees tend to reduce the prediction variance by aggregating many versions of the training data generated by k different

CART decision trees [Breiman, 1996].

Generating bootstrap samples introduces a random component into the tree building process, which induces a distribution of trees, and therefore a distribution of predicted values for each sample. However, the trees in bagging are not completely independent of each other since all of the original predictors are considered at every split of every tree. Considering many original samples and a relationship between predictors and response that a tree can model, then trees from different bootstrap samples may have similar structures to each other due to the underlying relationship. This characteristic (tree correlation) prevents bagging from optimally, reducing the variance of the predicted values Breiman [2001]. Dietterich [2000] developed the idea of random split selection, where trees are built using a random subset of the top k predictors at each split in the tree. After evaluating this method, Breiman [2001] constructed a unified algorithm called Random Forests. Each model in the ensemble is then used to generate a prediction for a new sample, and these m predictions are averaged to give the forest's prediction. Since the algorithm randomly selects predictors at each split, tree correlation will necessarily be lessened. Random forests (RF) is an ensemble machine learning method based on the construction of multiple CART decision trees to achieve a better performance than a "single" tree model (for either regression or classification). The main underlying technique used in random forests is bootstrap aggregating (Bagging). RF estimates "simple" decision trees by resampling the dataset and the feature space, obtaining their predicted response. Random Forests algorithm's predicted value is the mode in case of classification or the average value of the k different decision trees in case of regression.

Gradient Boosting Machine (GBM) [Friedman, 2002] was originally developed for classification problems and later extended to regression. The basic principles of GBM are as follows: given a loss function (e.g., error measure for regression) and a learner (e.g., regression trees), the algorithm tries to find an additive model that minimizes the loss function. The algorithm is initialized with the best guess of the response (e.g., the mean of the response in regression). The residual is calculated, and a model is then fit to the residuals to minimize the loss function. The current model is added to the previous one, and the procedure continues for a user-specified number of iterations. Therefore, GBM has clear similarities to Random Forests since an ensemble of decision trees estimates the final prediction. However, the way the ensembles are constructed differs substantially between each method. In Random Forests, all trees are created independently and contribute equally to the final model. In GBM, however, new trees are dependent on past ones and contribute unequally

to the final model. Despite these differences, both Random Forests and GBM offer competitive predictive performance.

Linear Regression models [Graybill, 1976] are the most common type of model used in the length of stay literature, mainly for inference problems. The advantage of this type of model is the ease of understanding and implementation. However, because of the assumption of linear relationship with the response, this model's performance in LoS prediction problems is not reported as competitive [Peres et al., 2020]. We recommend to include the Linear Regression (LR) and the Generalized Linear Model (GLM) [Nelder and Wedderburn, 1972] with Negative Binomial distribution in prediction tests just to have a baseline model for comparison reasons. LR (or Ordinary Least Squares Linear Regression) aims to find the plane that minimizes the error measure between the observed and predicted response. LR chooses the parameters of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable in the training data and those predicted by the linear function [Graybill, 1976]. GLM model is a flexible generalization of LR that allows for response variables that have error distribution other than Normal distributed. This model generalizes LR by allowing the linear model to be related to the response variable via a link function and allowing the magnitude of each measurement's variance to be a function of its predicted value [Nelder and Wedderburn, 1972].

As stated in Section 3.1.3, we recommend splitting the dataset into training and testing sets, using random sampling with the 80%/20% ratio. With the training set, we suggest using 5-fold cross-validation to obtain the best model parameters, in which each model set is evaluated in one part of the dataset and the other four are used to estimate the model. The hyperparameters of the model must be selected in order to minimize the prediction error, which should be calculated by the Root Mean Square Error (RMSE). We recommend testing the following parameters in each model [Caetano et al., 2014; Houthooft et al., 2015; Kuhn and Johnson, 2013]:

- SVR with radial kernel: $\sigma \in \{2^{-3}, 2^{-2}, ..., 2^{3}\}$; and $C \in \{2^{-3}, 2^{-2}, ..., 2^{3}\}$;
- SVR with linear kernel: $C \in \{2^{-15}, 2^{-14}, ..., 2^{3}\}$;
- kNN: $k \in \{2, 3, ..., 60\}$;
- GBM: $Interaction\ depth \in \{5, 10, 15, 20\}$; $Number\ of\ trees = 300$; $Shrinkage \in \{0.1, 0.01, 0.001\}$; $Minimal\ number\ of\ observations\ in\ each\ node = 20$.
- CART: $cp \in \{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$

- RF: *Number of variables to possibly split at each node* ($mtry$) $\in$ {5, 6, ..., 10}; *Minimal node size* $\in$ {5, 6, ..., 10}; *Splitting rule* $\in$ $\{variance, extratrees, maxstat, beta\}$;
- GLM with Negative Binomial distribution: *Link* $\in$ {*log*,*sqrt*,*identity*};
- Linear Regression and Bagging do not have tuning parameters.

After training the models and selecting the best set of hyper-parameters, we suggest comparing them using the following performance indicators: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2). It's important to analyze their calibration plotting predicted versus observed length of stay for the best models. Moreover, the performance indicators can be analyzed in LoS ranges, such as [0-3];]3-7];]7-10];]10-14]; and >14 days.

3.3 Predicting the Risk of Prolonged Stay

Another relevant information used by ICU managers is the risk of a patient being a prolonged stay, which makes possible to early identify prolonged stay patients and drive immediate quality improvement [Marik and Hedman, 2000; Rapoport et al., 2003; Verburg et al., 2014]. The response feature in this model is whether the patient presented a low or high ICU LoS. We proposed two different alternatives of prediction models for this problem, which can be used by managers depending on their objectives, as follows:

- The first alternative predicts a patient's risk to stay over 14 days, following the literature [Houthooft et al., 2015; Laupland et al., 2006; Zampieri et al., 2014]. This model aims to identify the prolonged stay patients regardless of their diagnosis at admission, which is an important tool to plan the resources and operations required to fulfill ICU care.
- The second proposition predicts a patient's risk to stay over a specific threshold, which is defined as the 90% percentile of ICU stay for his diagnosis group. This model aims to predict patients with critical conditions inside their diagnosis group to drive ICU care's immediate improvement.

Several literature papers defined the prolonged stay threshold based on clinical judgments [Azari et al., 2012; Hachesu et al., 2013; Houthooft et al., 2015; Laupland et al., 2006; Liu et al., 2006; Weissman et al., 2018; Zampieri et al., 2014]. Most of them defined prolonged stay as an ICU LoS bigger than 14 days [Houthooft et al., 2015; Laupland et al., 2006; Zampieri

et al., 2014]. We proposed model A following the literature, which predicts the patient's risk to stay over 14 days. This information is an important tool to plan the resources required. Moreover, we noted from the literature review performed in Chapter 2 that several papers noted a significant association between the reasons (or diagnosis) of ICU admission and the ICU length of stay [Peres et al., 2020]. Sepsis was the most analyzed factor, and most studies found that sepsis patients tend to have greater ICU stay [da Silva et al., 2015; Knaus et al., 1993; Makrygiannis et al., 2018; Morello et al., 2019]. Moreover, myocardial infarction, intracerebral hemorrhage, pulmonary edema, and subarachnoid hemorrhage were also found to be positively related to ICU LoS [Al Tehewy et al., 2010; Knaus et al., 1993; Kramer and Zimmerman, 2010; Lim et al., 2010]. Therefore, it may not be reasonable to use just one threshold to define prolonged stay for all patients since the expected length of stay may depend on the admission diagnosis. For this reason, we included the second alternative (model B), predicting the risk using a variable threshold defined as a stay over the 90% percentile of the patient's diagnosis group. To the best of our knowledge, no literature article noted this relevant behavior, giving more information to the ICU managers about patients with critical conditions inside their diagnosis group and improving ICU care.

We will train a classification model to predict the probability to be a prolonged stay. As our response variable is binary, some model training parameters must be changed. A standard performance metric used to optimize models is the Brier Score, which simultaneously addresses calibration, statistical consistency between the predicted probability and the observations, as well as sharpness [Gneiting and Raftery, 2007; Rufibach, 2010]. Equation 3-3 shows the Brier Score formulation. The terms x_i, p_i and n represent, respectively, the observed type of stay (0 or 1), the predicted probability of prolonged stay, and the number of observations. Still, other indicators should be analyzed, like the Area Under the Curve (AUC), the Positive Predictive Value (PPV), and the Negative Predictive Value (NPV). We will train the model to predict the risk of prolonged ICU LoS using the best-performing type of model and set of features obtained from the analysis of Section 3.2.

$$BrierScore = \frac{\sum_{i=1}^{n}(x_i - p_i)^2}{n} \tag{3-3}$$

3.4 Performing a Benchmarking Analysis between ICUs

The ICU length of stay can be a surrogate of cost and efficiency and typically reflects several aspects of care, including admission and discharge

policies, adherence to best practices, and patient safety. Insightful information can be obtained when this indicator is analyzed in association with data on ICU staffing and resources, bed availability and capacity strain, case mix, mortality and infection rates, and hospital structure [Salluh et al., 2017; Verburg et al., 2018a].

Rothen et al. [2007] evaluated ICU efficiency using the Standardized Resource Use (SRU). This measure uses the ICU length of stay to estimate the average amount of resources used per surviving patient in a specific unit. Equation 3-4 shows the SRU formulation, a measure of efficiency for each Unit (u). The terms $ObservedICULOS_{iu}$ and $ExpectedICULOS_{iu}$ represent the observed ICU LoS and the expected ICU LoS for each patient i attended in the referred ICU u. Rothen et al. [2007] estimated the expected ICU LoS using deciles of the SAPS3 severity score. In the first step, the aggregate data were stratified according to SAPS 3 admission score. For each stratum, the sum of ICU LoS of all patients in that stratum was calculated. For each stratum, this sum was divided by the number of patients, resulting in the expected ICU LoS. Therefore, to find the expected ICU LoS for each patient, one would have to get the patient admission SAPS3, observe in which stratum this SAPS3 is allocated, and get the referred expected ICU LoS.

$$SRU_u = \frac{\sum_{i=1}^{n} ObservedICULOS_{iu}}{\sum_{i=1}^{n} ExpectedICULOS_{iu}} \tag{3-4}$$

This SRU measure proposed by Rothen et al. [2007] is simple to be implemented and used by several managers and researchers to evaluate ICUs' efficiency [Bastos et al., 2020; Soares et al., 2015; Vincent et al., 2012; Wortel et al., 2021]. However, there is a limitation regarding how accurate is a prediction for ICU LoS using just the SAPS3 feature. We will demonstrate in Section 4.2.4 that the prediction accuracy using only severity score features tends to be inferior to using the main important features for ICU LoS prediction (selected by feature selection techniques). For that reason, Verburg et al. [2018a] proposed another way to measure ICU efficiency, named Standardized Length of Stay Ratio (SLOSR). SLOSR has a similar formulation compared to SRU (3-5). However, instead of using the SAPS3 to predict the expected ICU LoS, the authors build a prediction model using ordinary least square regression with a log-link function. SLOSR tends to improve case-mix correction, which is essential when comparing the LoS between ICUs (benchmarking) [Marik and Hedman, 2000; Rapoport et al., 2003; Verburg et al., 2017]. We used a similar approach and developed a SLOSR measure using our proposed prediction model of Section 3.2. This approach aims to perform a non-biased benchmarking analysis between ICUs. Since our model considers the patient's

main clinical features, the predictions tend to be case-mix corrected. Efficient ICUs have SLOSR lower than one (total observed < total expected LoS), while inefficient ICUs present SLOSR bigger than one (total observed > total expected LoS). Moreover, the lower the SLOSR, the more efficient the ICU is considered in terms of resource use.

$$SLOSR_u = \frac{\sum_{i=1}^{n} ObservedICULOS_{iu}}{\sum_{i=1}^{n} ExpectedICULOS_{iu}} \tag{3-5}$$

First, we will compare the grouped length of stay per Unit by plotting the sum of predicted and observed LoS for each ICU, showing the determination coefficient (R^2). This plot analyzes the calibration of our model to predict the grouped ICU LoS per Unit. We proposed another calibration measure, named overall SLOSR (see $SLOSR_T$ in Equation 3-6), which has the same formulation of SLOSR, but considers all patients and not just the patients from a specific ICU. Overall SLOSR bigger than one means that the model underestimates the predictions, while lower than one implies overestimation. We recommend future papers reporting this calibration measure, which is currently neglected by the literature. Regarding the particular SLOSR ($SLOSR_u$), we will use this measure to evaluate ICUs efficiency.

$$SLOSR_T = \frac{\sum_{i=1}^{n} ObservedICULOS_{i}}{\sum_{i=1}^{n} ExpectedICULOS_{i}} \tag{3-6}$$

Funnel plots can be used to present the values of a quality indicator associated with individual ICUs and compare these values to the benchmark. An example of a funnel plot is presented in Figure 3.4. The value of each unit's quality measure is plotted against a measure of its precision, often the number of admissions. Control limits, illustrated in dashed lines, indicate a range in which the quality measure's values would be expected to fall. The control limits form a "funnel" shape around the benchmark, presented as a horizontal line. If an ICU falls outside the control limits, it is seen as performing differently than expected, given the value of the benchmark [Mayer et al., 2009; Rakow et al., 2015; Spiegelhalter, 2005; Verburg et al., 2018b]. Incorrectly constructed funnel plots could lead to incorrect judgments being made about ICUs, which could represent severe consequences, especially if one use them to judge or choose ICUs. It is important to assume that ICUs inside the control limits perform according to the benchmark, while ICUs falling outside do not perform as expected [Verburg et al., 2018b]. In our example, ICUs 2, 5, 6, 9, 12, and 13 falls outside the 95% control limits, which indicates that these ICUs perform differently than the benchmark considering the confidence level. If we consider the 99.8% control limits, only ICUs 2, 5, and 9 would fall outside. We can also note that the overall SLOSR was close to one, showing an accurate model

calibration. Therefore, we used the funnel plot to analyze the calibration of our proposed SLOSR measure.

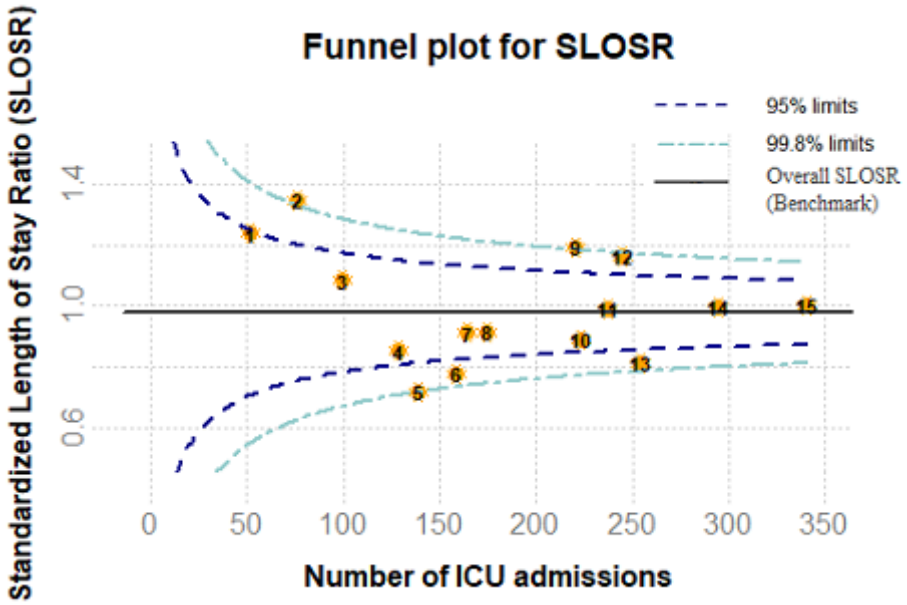


Figure 3.4: Example of funnel plot for the Standardized Length of Stay Ratio (SLOSR).

4
Application in a dataset with 109 mixed-type ICUs

This chapter will show the application of our proposed data-driven methodology to predict ICU LoS to a big Brazilian dataset of mixed-type ICUs. All analyses were performed using R software version 3.6.3.

4.1
Materials

This section presents the dataset that will be studied. The data represents a set of 109 mixed-type ICUs from 38 different Brazilian hospitals.

4.1.1
Inclusion Criteria

The extracted dataset contains a total of 103,195 independent admissions from January 01 to December 31, 2019. The complete dataset is a join of five tables of the hospital database: demographic and admission data, comorbidities, ICU complications (first 24 hours), physiological and laboratory data (first one hour), and secondary diagnosis. The inclusion criteria were as follows: patients aged 16 years old or more, with ICU LoS bigger than six hours, that presented previous hospital LoS lower than 60 days, with unit admission date bigger than hospital admission date, and presenting the main admission code. These criteria was defined following the literature [Peres et al., 2020; Verburg et al., 2017] and clinical judgments. After applying the criteria, the final dataset remained with 99,492 admissions, as illustrated in Figure 4.1.

4.1.2
Features

Table 4.1 presents the complete dataset dictionary. The dataset presets more than 100 features, which shows the complexity of our analysis.

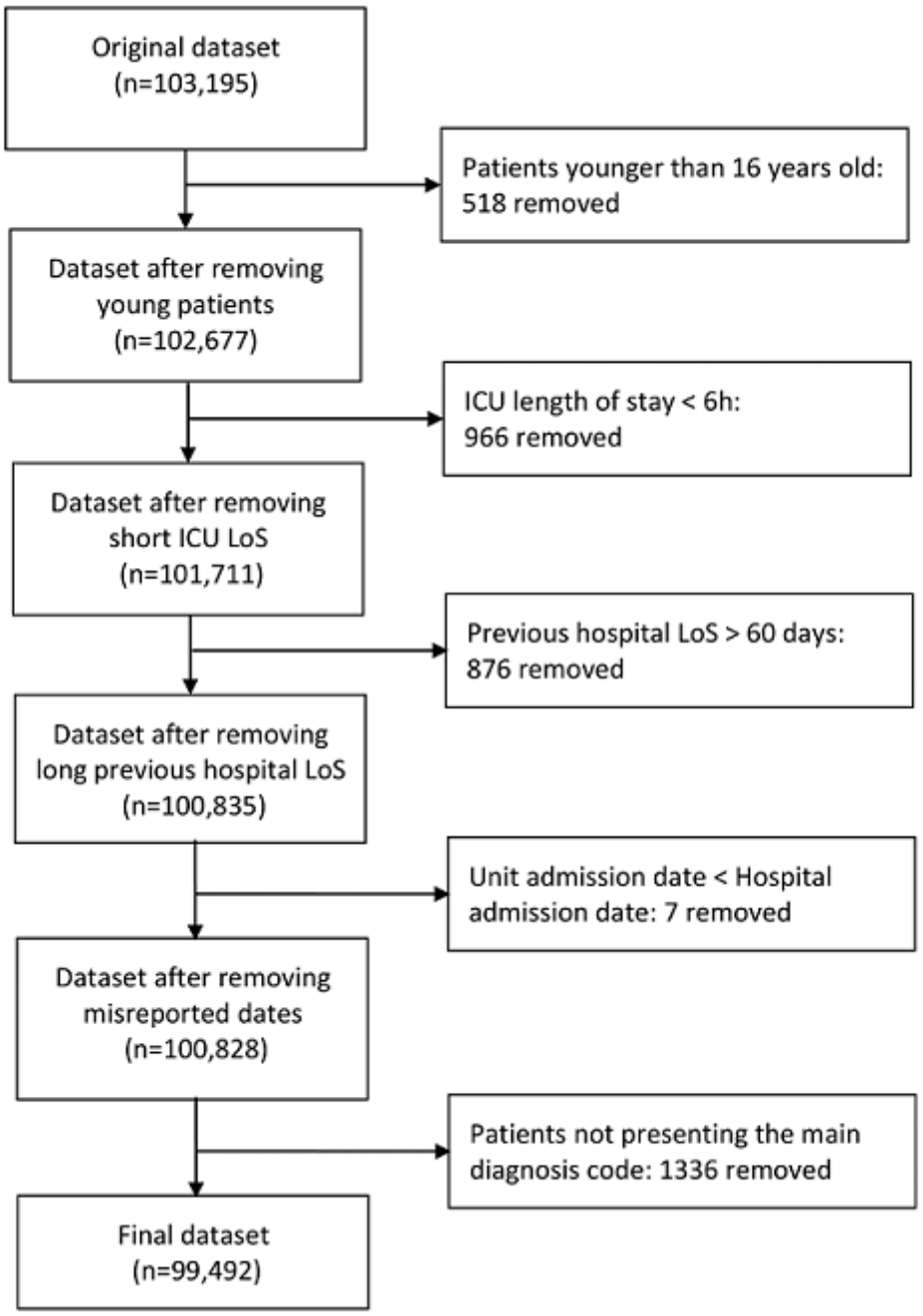


Figure 4.1: Registries inclusion criteria

Table	Features	Description	Type	Missing (%)
Admission	IsHospitalReadmission	If the patient is a readmission to the hospital	Binary	0%
Admission	Age	Age	Numerical	0%
Admission	Gender	Gender	Categorical	0%
Admission	BMI	Body Mass Index	Numeric	41%
Admission	IsReadmission	If the patient is a readmission to the ICU	Binary	0%
Admission	IsReadmission24h	If the patient is a readmission to the ICU in the last 24 hours	Binary	0%
Admission	IsReadmission48h	If the patient is a readmission to the ICU in the last 48 hours	Binary	0%
Admission	ICDCode	International Classification of Diseases (ICD) codes	Categorical	100%
Admission	LoSPriorUnitAdmission	The hospital length of stay prior to the admission in the ICU	Numerical	0%
Admission	UnitLengthStay	The patient length of stay in the ICU (outcome)	Numerical	0%
Admission	AdmissionTypeName	Type of ICU admission	Categorical	0%
Admission	AdmissionSourceName	Source of ICU admission	Categorical	1%
Admission	CharlsonComorbidityIndex	Charlson Comorbidity Index	Numerical	0%
Admission	MFIpoints	Modified Frailty Index points	Numerical	2%
Admission	MFIScore	Modified Frailty Index score	Numerical	1%
Admission	FrailPatientMFI	Frail patient (yes or no)	Binary	1%
Admission	Saps3Points	Simplified Acute Physiology Score 3 points	Numerical	0%
Admission	Saps3DeathProbabilityEquation	Simplified Acute Physiology Score 3 probability	Numerical	0%
Admission	SofaScore	Sequential Organ Failure Assessment score	Numerical	0%
Admission	AdmissionMainDiagnosisName	Main diagnosis of ICU admission	Categorical	30%
Diagnoses	SecondaryDiagnosis	Secondary diagnoses of ICU admission	Categorical	0%
Comorbidities	ChronicHealthStatusName	Name of Chronic Health status	Categorical	1%
Comorbidities	IsChfNyhaClass23	Cardiac Heart Failure class II or III of NYHA Classification	Binary	1%
Comorbidities	IsChfNyhaClass4	Cardiac Heart Failure class IV of NYHA Classification	Binary	1%
Comorbidities	IsCrfNoDialysis	Chronic Renal Failure and dialysis was not required	Binary	1%
Comorbidities	IsCrfDialysis	Chronic Renal Failure and dialysis was required	Binary	1%
Comorbidities	IsCirrhosisChildAB	Child-Pugh class "A" or "B" for Chronic Liver Disease	Binary	1%
Comorbidities	IsCirrhosisChildC	Child-Pugh class "C" for Chronic Liver Disease	Binary	1%
Comorbidities	IsHepaticFailure	Hepatic Failure	Binary	1%
Comorbidities	IsSolidTumorLocoregional	Solid Tumor Locoregional	Binary	1%
Comorbidities	IsSolidTumorMetastatic	Solid Tumor Metastatic	Binary	1%
Comorbidities	IsHematologicalMalignancy	Hematological Malignancy	Binary	1%
Comorbidities	IsImmunossupression	Immunossupression	Binary	1%
Comorbidities	IsSevereCopd	Severe COPD	Binary	1%
Comorbidities	IsSteroidsUse	Steroids Use	Binary	1%
Comorbidities	IsAids	Aids	Binary	1%
Comorbidities	IsArterialHypertension	Arterial Hypertension	Binary	1%
Comorbidities	IsAsthma	Asthma	Binary	1%
Comorbidities	IsDiabetesUncomplicated	Diabetes Uncomplicated	Binary	1%
Comorbidities	IsDiabetesComplicated	Diabetes Complicated	Binary	1%
Comorbidities	IsAngina	Angina	Binary	1%
Comorbidities	IsPreviousMI	Previous Myocardial Infarction	Binary	1%
Comorbidities	IsCardiacArrhythmia	Cardiac Arrhythmia	Binary	1%
Comorbidities	IsDeepVenousThrombosis	Deep Venous Thrombosis	Binary	1%
Comorbidities	IsPeripheralArteryDisease	Peripheral Artery Disease	Binary	1%
Comorbidities	IsChronicAtrialFibrilation	Chronic Atrial Fibrilation	Binary	1%
Comorbidities	IsRheumaticDisease	Rheumatic Disease	Binary	1%
Comorbidities	IsStrokeSequelae	Stroke with sequelae	Binary	1%
Comorbidities	IsStrokeNoSequelae	Stroke without sequelae	Binary	1%
Comorbidities	IsDementia	Dementia	Binary	1%
Comorbidities	IsTobaccoConsumption	Tobacco Consumption	Binary	1%
Comorbidities	IsAlcoholism	Alcoholm	Binary	1%
Comorbidities	IsPsychiatricDisease	Psychiatric Disease	Binary	1%
Comorbidities	IsMorbidObesity	Morbid Obesity	Binary	1%
Comorbidities	IsMalnourishment	Malnourhment	Binary	1%
Comorbidities	IsPepticDisease	Peptic Disease	Binary	1%
Comorbidities	IsSolidOrganTransplant	Solid Organ Transplant	Binary	1%

Table	Features	Description	Type	Missing (%)
Comorbidities	IsAutologousBMT	Autologous Bone Marrow Transplant	Binary	1%
Comorbidities	IsAllogeneicBMT	Allogeneic Bone Marrow Transplant	Binary	1%
Comorbidities	IsOtherSolidOrganTransplant	Other Solid Organ Transplant	Binary	1%
Comorbidities	IsCardiacTransplant	Cardiac Transplant	Binary	1%
Comorbidities	IsCombinedLiverkidneyTransplant	Combined Liver kidney Transplant	Binary	1%
Comorbidities	IsCombinedPancreaskidneyTransplant	Combined Pancreas kidney Transplant	Binary	1%
Comorbidities	IsLiverTransplant	Liver Transplant	Binary	1%
Comorbidities	IsIntestinalTransplant	Intestinal Transplant	Binary	1%
Comorbidities	IsPancreasTransplant	Pancreas Transplant	Binary	1%
Comorbidities	IsLungTransplant	Lung Transplant	Binary	1%
Comorbidities	IsKidneyTransplant	Kidney Transplant	Binary	1%
Comorbidities	IsHypothyroidism	Hypothyroidism	Binary	1%
Comorbidities	IsHyperthyroidism	Hyperthyroidism	Binary	1%
Comorbidities	IsDyslipidemias	Dyslipidemias	Binary	1%
Comorbidities	IsChemotherapy	Chemotherapy	Binary	1%
Comorbidities	IsRadiationTherapy	Radiation Therapy	Binary	1%
Comorbidities	IsHistoryOfPneumonia	History of pneumonia	Binary	1%
Complications	IsRespiratoryFailure	Respiratory Failure	Binary	2%
Complications	IsMechanicalVentilation	Mechanical Ventilation	Binary	2%
Complications	IsNonInvasiveVentilation	Non-invasive Ventilation	Binary	2%
Complications	IsVasopressors	Vasopressors	Binary	2%
Complications	IsCardiacArrhythmias	Cardiac Arrhythmias	Binary	2%
Complications	IsCardiopulmonaryArrest	Cardio pulmonary Arrest	Binary	2%
Complications	IsAcuteKidneyInjury	Acute Kidney Injury	Binary	2%
Complications	IsRenalReplacementTherapy	Renal Replacement Therapy	Binary	2%
Complications	IsGastrointestinalBleeding	Gastrointestinal Bleeding	Binary	2%
Complications	IsIntracranialMassEffect	Intracranial Mass Effect	Binary	2%
Complications	IsNeutropenia	Neutropenia	Binary	2%
Complications	IsAsystole	Asystole	Binary	2%
Complications	IsPulselessElectricalActivity	Pulseless Electrical Activity	Binary	2%
Complications	IsVentricularSustainedCardiopulmonary	Ventricular Sustained Cardiopulmonary	Binary	2%
Complications	IsAcuteAtrialFibrilation	Acute Atrial Fibrilation	Binary	2%
Complications	IsAtrialFlutter	Atrial Flutter	Binary	2%
Complications	IsVentricularSustainedArrhythmia	Ventricular Sustained Arrhythmia	Binary	2%
Laboratory	LowestSystolicBloodPressure1h	Lowest Systolic Blood Pressure (first 1 hour of admission)	Numerical	3%
Laboratory	LowestDiastolicBloodPressure1h	Lowest Diastolic Blood Pressure (first 1 hour of admission)	Numerical	3%
Laboratory	LowestMeanArterialPressure1h	Lowest Mean Arterial Pressure (first 1 hour of admission)	Numerical	3%
Laboratory	LowestGlasgowComaScale1h	Lowest Glasgow Coma Scale (first 1 hour of admission)	Numerical	10%
Laboratory	LowestPlateletsCount1h	Lowest Platelets Count (first 1 hour of admission)	Numerical	14%
Laboratory	LowestPH1h	Lowest PH (first 1 hour of admission)	Numerical	94%
Laboratory	LowestPaO21h	Lowest PaO2 (first 1 hour of admission)	Numerical	94%
Laboratory	LowestPaCO21h	Lowest PaCO2 (first 1 hour of admission)	Numerical	94%
Laboratory	LowestFiO21h	Lowest FiO2 (first 1 hour of admission)	Numerical	92%
Laboratory	LowestPaO2FiO21h	Lowest PaO2FiO2 (first 1 hour of admission)	Numerical	95%
Laboratory	HighestHeartRate1h	Highest Heart Rate (first 1 hour of admission)	Numerical	3%
Laboratory	HighestRespiratoryRate1h	Highest Respiratory Rate (first 1 hour of admission)	Numerical	4%
Laboratory	HighestTemperature1h	Highest Temperature (first 1 hour of admission)	Numerical	4%
Laboratory	HighestLeukocyteCount1h	Highest LeukocyteCount (first 1 hour of admission)	Numerical	14%
Laboratory	HighestCreatinine1h	Highest Creatinine (first 1 hour of admission)	Numerical	14%
Laboratory	HighestBilirubin1h	Highest Bilirubin (first 1 hour of admission)	Numerical	58%
Laboratory	HighestPH1h	Highest PH (first 1 hour of admission)	Numerical	75%
Laboratory	HighestPaO21h	Highest PaO2 (first 1 hour of admission)	Numerical	76%
Laboratory	HighestPaCO21h	Highest PaCO2 (first 1 hour of admission)	Numerical	76%
Laboratory	HighestFiO21h	Highest FiO2 (first 1 hour of admission)	Numerical	61%
Laboratory	HighestPaO2FiO21h	Highest PaO2FiO2 (first 1 hour of admission)	Numerical	84%
Laboratory	HighestArterialLactate1h	Highest Arterial Lactate (first 1 hour of admission)	Numerical	71%
Laboratory	Urea	Urea	Numerical	16%
Laboratory	BUN	BUN	Numerical	16%

Table 4.1: Complete data dictionary

The basic descriptive analysis for some numeric features is presented in Table 4.2. The average ICU length of stay was 4.54 with a standard deviation of 4.82. The patients were on average 61 years old, with BMI equal to 26.8, previous length of stay before ICU admission equal to 1.8 days, with on average 42.7 SAPS3 points, and 1.16 SOFA score. Table 4.3 presents the description for some categorical features. 52.8% of the patients were female gender, 81.24% were admitted for clinical reasons, 68.61% came from the emergency, 2.7% were readmissions, 4.3% needed mechanical ventilation, 5.8% needed non-invasive ventilation, and 5.2% used a vasopressor. The ICU length of stay was more considerable for clinical or urgent patients coming from the ward, semi-intensive unit, another unit, or another hospital. The readmissions and the use of invasive and non-invasive supports seem to be important features, which increased approximately two times the ICU length of stay. In Section 4.2.2, we will explore more details about the association of the covariates with the ICU LoS, and in Section 4.2.3, we will perform a collinearity analysis of the variables.

Feature	Mean	SD
UnitLengthStay	4.54	4.82
Age	61.57	20.73
BMI	26.85	5.62
LoSPriorUnitAdmission	1.82	6.05
Saps3Points	42.71	12.93
SofaScore	1.16	2.02
Glasgow	14.42	1.91
Urea	51.44	58.37
LeukocyteCount	11.77	18.70
Creatinine	1.22	1.62
Bilirubin	0.61	1.45
HeartRate	81.16	18.91
RespiratoryRate	18.70	5.14
Temperature	36.15	0.64

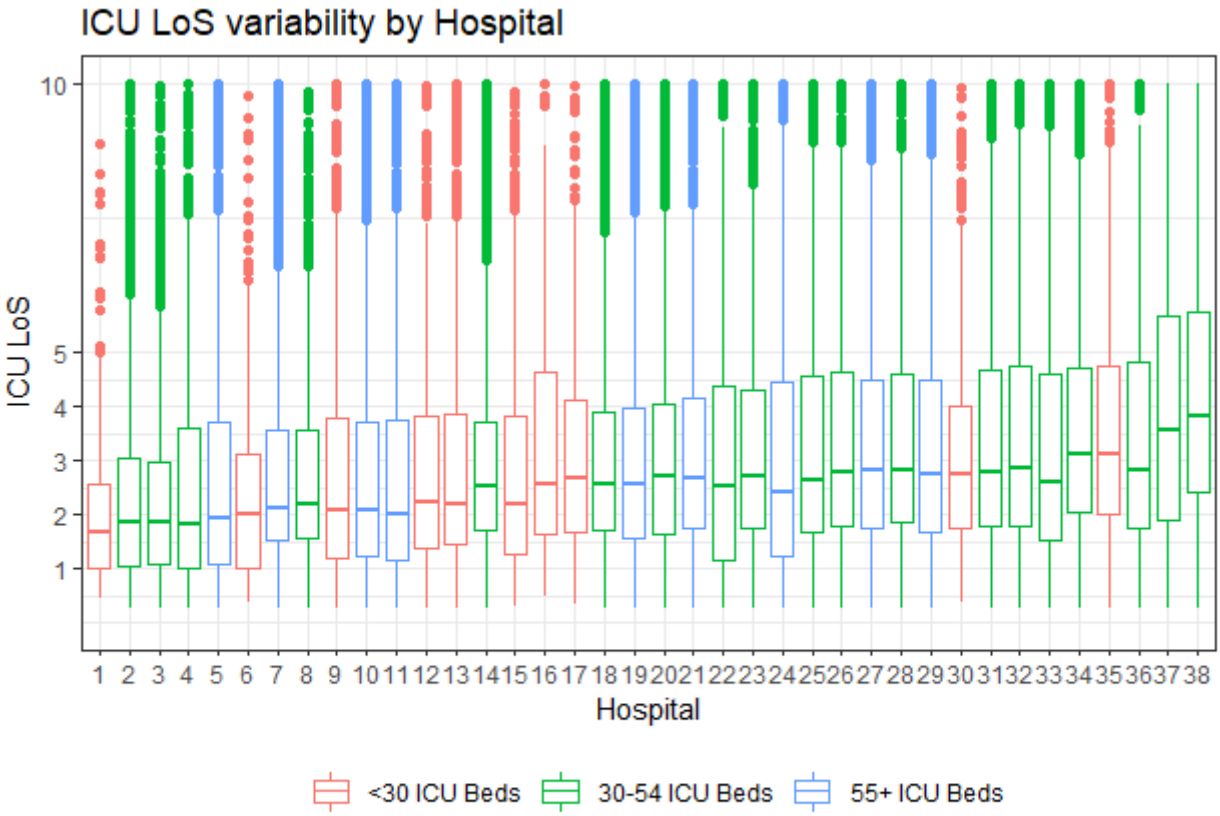
Table 4.2: Descriptive analysis for numeric features

Features		Mean ICU LoS	SD	n	%
Gender					
	Female	4.51	4.76	52517	52.77%
	Male	4.55	4.87	47000	47.22%
	Undefined	8.21	6.21	10	0.01%
Admission Type					
	Clinical	4.86	4.93	80852	81.24%
	Elective surgery	2.77	3.56	14951	15.02%
	Urgent surgery	4.38	5.01	3724	3.74%
Admission Source					
	Emergency	4.41	4.46	68281	68.61%
	Surgical center	3.13	3.99	13935	14.00%
	Ward	7.04	6.08	5987	6.02%
	Other ICU	7.58	6.80	4271	4.29%
	Hemodynamics room	2.68	3.23	3998	4.02%
	Transfers	6.18	6.18	1617	1.62%
	Semi-intensive Unit	7.93	6.69	658	0.66%
	Others	2.43	3.60	780	0.78%
Readmission					
	No	4.46	4.74	96806	97.27%
	Yes	7.01	6.34	2721	2.73%
Mechanical Ventilation					
	No	4.24	4.44	95256	95.71%
	Yes	10.90	7.60	4271	4.29%
Non Invasive Ventilation					
	No	4.36	4.66	93750	94.20%
	Yes	7.20	6.26	5777	5.80%
Vasopressors					
	No	4.26	4.50	94317	94.77%
	Yes	9.32	7.18	5210	5.23%

Table 4.3: Descriptive analysis for categorical features

4.1.3
Hospitals' Descriptive Analysis

The dataset includes 38 hospitals with different ICU types and patient case mixes. The distribution of hospitals' size in terms of number of ICU beds was as follows: minimum = 7 ICU beds; first quarter = 29; median = 47; third quarter = 54; maximum = 110. Figure 4.2 presents the variability of ICU length of stay per hospital ordered by median ICU LoS. We separated the analysis by hospital size: red color represents hospitals with less than 30 ICU beds, green color for hospitals with 30 to 54 beds, and blue color for hospitals with more than 55 beds. We can see in general that the hospitals are not homogeneous, presenting different distributions of ICU LoS. The hospital with the lowest median ICU LoS was "Hospital 8" (1.7 days), and the one with the highest median LoS was "Hospital 17" (4.7 days). Figure 4.3 shows the same analysis but grouping the hospitals of the same size. We can note that medium-size hospitals presented a bigger ICU LoS than other sizes (2.69 days for small; 2.85 days for medium; 2.75 days for big). However, this difference was small and may be related to the case-mix attended by each ICU.



Hospital	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
Median LoS	1.7	1.9	2.0	2.0	2.2	2.2	2.3	2.3	2.4	2.5	2.5	2.5	2.6	2.6	2.6	2.7	2.8	2.8	2.8	2.8	2.9	2.9	2.9	2.9	3.0	3.0	3.0	3.0	3.0	3.1	3.1	3.1	3.2	3.5	3.7	3.8	3.9	4.7

Figure 4.2: Boxplot of ICU LoS by Hospitals

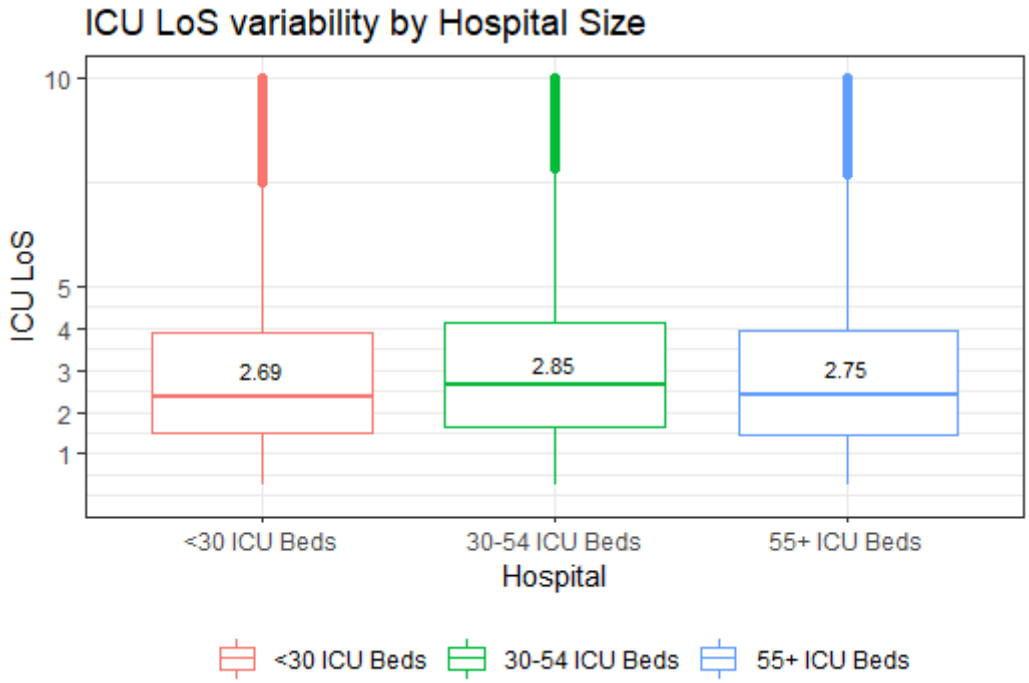


Figure 4.3: Boxplot of ICU LoS by Hospital Size

We had the following description regarding the types of ICUs: 70 general (mixed medical and surgical), 25 cardiac, five surgical, four neurological, three oncological, and two orthopedic-type ICUs. The number of admissions in each type was 69973, 20284, 5050, 1791, 1474, and 1174, and the average ICU LoS was 4.9, 5.1, 3.4, 3.8, 4.7, and 8.02, respectively. The distribution of length of stay for each type of ICU is presented in Figure 4.4. General, cardiac, and oncological ICUs presented a similar behavior. Orthopedic ICUs showed a significantly higher ICU LoS, while surgical and neurological-type presented a lower length of stay. This figure indicates that our database includes a heterogeneity set of ICUs, with different distributions of ICU LoS.

The next sections will present the results obtained after applying the data-driven methodology to predict ICU length of stay (presented in Section 3) to our dataset.

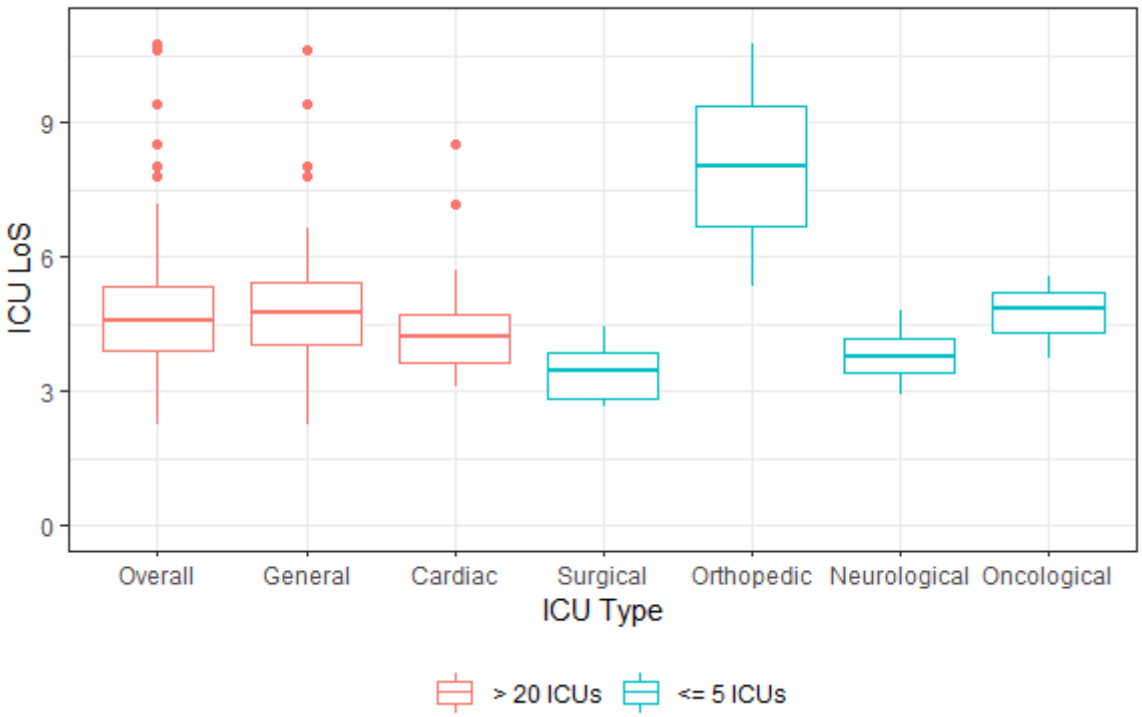


Figure 4.4: Boxplot of ICU LoS for each ICU type

4.2 Data Preparation and Preprocessing

First, we will show the results of the application of data preparation and preprocessing methodologies, as presented in Section 3.1.

4.2.1 Data Preparation

We analyzed the extracted dataset and adjusted issues that occurred for some features, as follows:

- Feature "Hospital length of stay prior to unit admission" had some missings values. So, we used the dates and updated the calculation of this feature ("Unit Admission Date" - "Hospital Admission Date").
- Feature "Gender": "Undefined" was replaced by “Not informed” and then will be imputed.
- Feature "Admission source" was reclassified from eight categories to the following three: "Surgical center", "Ward/Room/Semi-intensive Unit", and "Home-care/Transfers/Others".

Regarding the feature engineering, we proposed the following new features from existing ones: number of first day complications (“n_complication”), and presence of any first day complication (“has_complication”).

4.2.2 Visualization and Data Cleaning

As illustrated in Table 4.1, we analyzed the behavior of the missing data for each variable. Following White et al. [2011], features with more than 30% of missing were excluded from the analysis: "ICDCode" (100%); "LowestPaCO21h" (94%); "LowestPaO21h" (94%); "LowestFiO21h" (92%); "HighestPaO21h" (75%); "HighestPaCO21h" (75%); and "HighestFiO21h" (61%). The following features were not removed in this step because of their clinical relevance: PaO2FiO2, PH, Lactate, Bilirubin and BMI. Variables with less than 30% of missing will be imputed.

Regarding the descriptive statistical analysis, we applied a univariate analysis between the explanatory variables and the ICU LoS. For numerical variables we used Pearson Correlation, and for categorical ones we used Cramer's V.

Table 4.4 shows the Pearson Correlation between the numeric variables and ICU LoS. SAPS3, Glasgow and SOFA presented higher correlation with

Feature	Correlation
Saps3DeathProbabilityStandardEquation	0.14
Saps3Points	0.14
LowestGlasgowComaScale1h	-0.13
SofaScore	0.12
LengthHospitalStayPriorUnitAdmission	0.09
n_complication	0.09
MFIpoints	0.09
MFIScore	0.09
Age	0.07
HighestRespiratoryRate1h	0.07
CharlsonComorbidityIndex	0.06
HighestHeartRate1h	0.05
BUN	0.04
Urea	0.04
PaO2FiO2	0.02
HighestTemperature1h	0.02
HighestCreatinine1h	0.02
LowestDiastolicBloodPressure1h	-0.02
HighestLeukocyteCount1h	0.01
PH	0.01
LowestPlateletsCount1h	0.01
BMI	-0.01
Bilirubin	0.01
LowestMeanArterialPressure1h	-0.01
Lactate	0.00
LowestSystolicBloodPressure1h	0.00

Table 4.4: Correlation with LoS for numeric variables

LoS (0.14, -0.13, and 0.12, respectively). We noted that some features may be correlated with each other (e.g., "Saps3DeathProbability" and "Saps3Points"), which will be treated in the preprocessing step.

Table 4.5 shows the Cramer's V coefficient between the categorical features and ICU LoS. We can observe that the admission main diagnosis, the use of mechanical ventilation, and the use of vasopressors has greater correlation with LoS (0.29, 0.26, and 0.22, respectively).

Feature	Correlation
AdmissionMainDiagnosisName	0.29
IsMechanicalVentilation	0.26
IsVasopressors	0.22
has_complication	0.22
AdmissionReasonName	0.20
AdmissionSourceName	0.17
IsRespiratoryFailure	0.17
IsDementia	0.16
FrailPatientMFI	0.15
IsNonInvasiveVentilation	0.13
IsArterialHypertension	0.13
IsSevereCopd	0.10
IsCrf	0.09
IsStroke	0.08
IsReadmission	0.08
Sec_Sepseechoqueséptico	0.08
IsChronicAtrialFibrilation	0.08
IsRenalReplacementTherapy	0.08
IsAcuteKidneyInjury	0.08
ChfNyha	0.07
IsHospitalReadmission	0.07
IsDiabetes	0.07
IsImmunossupression	0.06
Sec_Monitoraçãoobservaçãoclínica	0.06
Sec_Pneumoniacomunitária	0.05
TumorSolido	0.05
IsCardiacArrhythmia	0.05
IsDeepVenousThrombosis	0.05
IsMalnourishment	0.05
IsSteroidsUse	0.04

Table 4.5: Correlation with LoS for categorical variables

Regarding the data cleaning, we analyzed the boxplot of each numeric feature. 376 values presented extreme outliers and were replaced by "not informed" (Supplementary Table A.1). These values will be imputed in the preprocessing step. For ICU LoS, following statistical and clinical judgments, we did the truncation at 21 days. This value corresponds to the 96% percentile of ICU LoS distribution, as can be seen in Figure 4.5.

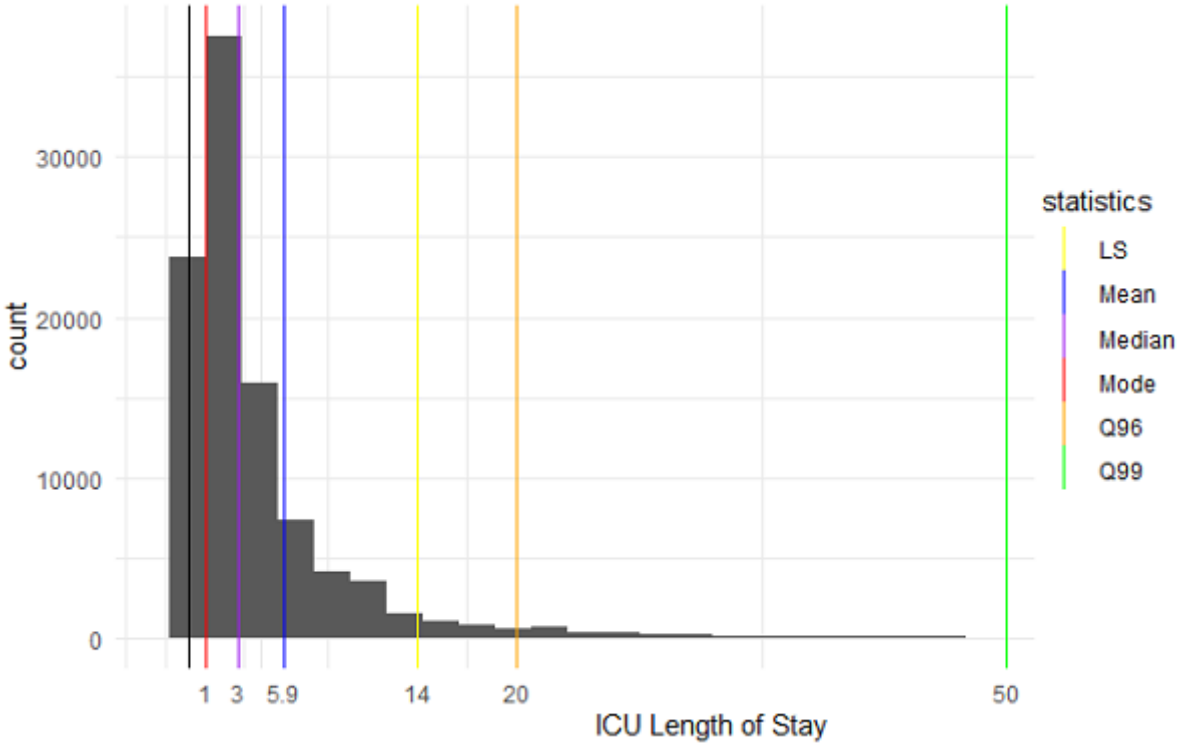


Figure 4.5: Histogram of ICU LoS

4.2.3 Data Preprocessing

In this section, we will present the results of data preprocessing, which includes: dimension reduction, imputation, feature selection, transformation to resolve skewness, normalization and one-hot encoding. Before preprocessing the dataset, we split the data into training and testing set using a 80%/20% proportion.

We applied two methods to reduce the dimension (complexity) of our dataset. First, we removed features with zero or near-zero variance. The following variables presented near-zero variance and were removed from the dataset: "IsReadmission24h", "IsReadmission48h", "IsOtherSolidOrganTransplant", "IsAtrialFlutter", "IsCombinedPancreaskidneyTransplant", "IsHyperthyroidism", "IsAllogeneicBMT", "IsAutologousBMT", "Is-

PepticDisease", "IsNeutropenia", "IsVentricularSustainedCardiopulmonary", "IsCombinedLiverkidneyTransplant", and "PaO2FiO2". Second, we did a collinearity analysis and removed the correlated features. Table 4.6 shows the collinearity analysis for numeric features, and Table 4.7 for categorical ones. For numeric variables, we used the recommended threshold of 0.75 to exclude collinear features [Kuhn and Johnson, 2013], and for categorical ones, we used the recommended threshold of 0.5 [Cohen, 1988]. The following features reported high correlation with others and were removed: "BUN", "MFIpoints", "Saps3DeathProbabilityStandardEquation", "LowestDiastolicBloodPressure1h", and "LowestSystolicBloodPressure1h" (for numeric features); and "IsCardiopulmonaryArrest", "IsImmunossupression", "AdmissionReasonName", and "has_complication" (for categorical features).

Feature 1	Feature 2	Correlation
BUN	Urea	1.00
MFIScore	MFIpoints	1.00
Saps3DeathProbabilityStandardEquation	Saps3Points	0.94
LowestMeanArterialPressure1h	LowestDiastolicBloodPressure1h	0.93
LowestMeanArterialPressure1h	LowestSystolicBloodPressure1h	0.89
LowestDiastolicBloodPressure1h	LowestSystolicBloodPressure1h	0.66
SofaScore	Saps3DeathProbabilityStandardEquation	0.64
Saps3Points	Age	0.63
LowestGlasgowComaScale1h	SofaScore	-0.60
SofaScore	Saps3Points	0.58
MFIpoints	Age	0.58
MFIScore	Age	0.58
n_complication	SofaScore	0.56
Saps3DeathProbabilityStandardEquation	Age	0.52
LowestGlasgowComaScale1h	Saps3DeathProbabilityStandardEquation	-0.52

Table 4.6: Collinearity analysis for numeric variables

Feature 1	Feature 2	Correlation
IsAsystole	IsCardiopulmonaryArrest	0.72
IsChemotherapy	IsImmunossupression	0.70
IsVasopressors	IsMechanicalVentilation	0.61
IsPulselessElectricalActivity	IsCardiopulmonaryArrest	0.53
has_complication	IsNonInvasiveVentilation	0.50
has_complication	IsRespiratoryFailure	0.48
has_complication	IsVasopressors	0.47
AdmissionSourceName	AdmissionReasonName	0.47
IsRadiationTherapy	IsChemotherapy	0.44
IsSteroidsUse	IsImmunossupression	0.43
has_complication	IsMechanicalVentilation	0.43
IsRadiationTherapy	IsImmunossupression	0.42
AdmissionMainDiagnosisName	AdmissionReasonName	0.42
IsCrf	IsRenalReplacementTherapy	0.42
TumorSolido	IsChemotherapy	0.42
TumorSolido	IsImmunossupression	0.41
TumorSolido	IsLungTransplant	0.41
TumorSolido	IsIntestinalTransplant	0.39
AdmissionSourceName	IsReadmission	0.37
IsAngina	FrailPatientMFI	0.37
IsMechanicalVentilation	IsRespiratoryFailure	0.37
has_complication	IsAcuteKidneyInjury	0.36
IsGastrointestinalBleeding	AdmissionReasonName	0.35
IsDiabetes	IsArterialHypertension	0.35
IsCardiacTransplant	IsSolidOrganTransplant	0.34
TumorSolido	IsLiverTransplant	0.34
IsDiabetes	FrailPatientMFI	0.34
IsMorbidObesity	AdmissionReasonName	0.34
ChfNyha	FrailPatientMFI	0.32
IsPreviousMI	FrailPatientMFI	0.32
IsVentricularSustainedArrhythmia	IsAcuteKidneyInjury	0.32
IsStroke	FrailPatientMFI	0.31
IsMechanicalVentilation	AdmissionReasonName	0.30
IsVentricularSustainedArrhythmia	IsSteroidsUse	0.30
IsCirrhosis	IsHepaticFailure	0.30
IsDementia	FrailPatientMFI	0.30
IsMalnourishment	IsSteroidsUse	0.29
Sec_Sepseechoqueséptico	AdmissionReasonName	0.29

Table 4.7: Collinearity analysis for categorical variables

After reducing the dimension of our dataset, we applied the MICE imputation algorithm to the variables with incomplete data. Therefore, the MICE imputation was applied to the following features: "BMI", "SofaScore", "Urea", "HighestCreatinine1h", "LowestPlateletsCount1h", "LowestGlasgowComaScale1h", "HighestTemperature1h", "HighestRespiratoryRate1h", "CharlsonComorbidityIndex", "LowestMeanArterialPressure1h", "HighestHeartRate1h", and "MFIScore". We also observed some laboratory variables not missing at random (MAR). These features were not collected because of the lack of need for these tests. In these cases, we considered the missing values as normal values for each variable, replacing by the median value. We applied this procedure to the following features: "PH", "Lactate" and "Bilirubin".

The next steps were the feature selection and the transformations to resolve skewness, which we will present in the next section. The preprocessing techniques related to normalization and one-hot encoding were applied depending on the requirement of each regression model.

4.2.4 Preprocessing Sensitivity Analysis

Before running all models to our dataset, we analyzed some scenarios of dataset transformations created by the preprocessing techniques to understand which one would best predict ICU LoS. These scenarios were modeled considering the GBM model. Therefore, the description of each sensitivity analysis is presented as follows:

– Analysis 1 - Comparison between different grouping strategies to the feature "Main Diagnosis": The original feature presents 851 diagnosis codes. We will test the original dataset and two alternatives of grouping this feature: into 19 classes and into eight classes. The grouping rule to this feature was based on the similarity of the diagnoses regarding the distribution of length of stay.

Table 4.8 shows the results for each grouping strategy. All three scenarios tested presented close results in RMSE, MAE, and R^2, both for training and testing. Therefore, we will select the scenario with less complexity (eight groups), desiring to improve the efficiency of the models.

Type of grouping for "Main Diagnosis"	Testing set RMSE	MAE	R^2	Cor	Training set RMSE	(SD)	MAE	(SD)	R^2	(SD)
Original dataset with 851 diagnosis codes	3.97	2.64	31%	0.554	3.96	0.06	2.60	0.03	33%	0.008
Dataset with 19 grouped diagnosis	3.97	2.64	31%	0.554	3.96	0.03	2.60	0.01	32%	0.005
Dataset with 8 grouped diagnosis	**3.99**	**2.65**	**30%**	**0.550**	**3.96**	**0.03**	**2.59**	**0.01**	**32%**	**0.005**

Table 4.8: Analysis of different grouping strategies to the feature "Main Diagnosis"

– Analysis 2 - Comparison between different types of transformation for the response variable "ICU LoS": We will test the ICU LoS in its original form, truncated at 21 days, and Box-Cox transformed.

Table 4.9 shows the results for each transformation strategy for ICU LoS. To make the comparison without any bias, we updated the predictions bigger than 21 days of all strategies truncating at 21 days. As can be seen from Table 4.9, the scenario with truncated LoS presented best results for all performance indicators (RMSE = 3.99; MAE = 2.65; R^2 = 30%; Correlation = 0.55). Therefore, we selected this strategy for ICU LoS transformation.

Type of LoS	Testing set RMSE	MAE	R^2	Cor
Truncated LoS	**3.99**	**2.65**	**30%**	**0.55**
Original LoS	4.05	2.69	28%	0.54
BoxCox LoS (Truncated)	4.09	2.50	27%	0.53
BoxCox LoS (Original)	4.15	2.61	13%	0.50

Table 4.9: Analysis of different transformations for the response variable "ICU LoS"

– Analysis 3 - Feature selection analyses: Comparison between different sets of features.

A relevant analysis included in this study was to test different sets of features with feature selection techniques. Severity scores like SAPS3, SOFA, MIF, and Charlson Comorbidity include several covariates in their formulation, and this correlation could be harmful to the model. First, we tested the influence of severity scores alone to see the explanation produced by these features. Then, we tested scenarios including and excluding severity scores features before applying the feature selection techniques. Finally, we tested scenarios excluding features with a high proportion of missing to test whether

their inclusion is effectively good to the model or not. Table 4.10 shows the results for scenarios with different sets of features. The scenarios represent the following characteristics, respectively:

- Baseline scenario: dataset including all features (no feature selection);
- Severity score analysis: dataset only including severity scores (no feature selection); and dataset excluding severity scores (no feature selection);
- Feature selection analyses: dataset applying feature selection using Treebag-RFE; dataset applying feature selection using RF-RFE; dataset excluding severity scores and applying feature selection using Treebag-RFE; dataset including the ICU code, excluding severity scores and applying feature selection using Treebag-RFE;
- Missing data analyses: dataset excluding imputed features that presented more than 15% of missing; dataset excluding imputed features that presented more than 3% of missing.

Severity scores, feature selection and missing data analyses	**Testing set**				**Training set**					
	RMSE	**MAE**	**R^2**	**Cor**	**RMSE**	**(SD)**	**MAE**	**(SD)**	**R^2**	**(SD)**
Dataset with all features (90 features)	3.99	2.65	30%	0.550	3.96	0.03	2.59	0.01	32%	0.005
Severity scores analysis										
Dataset considering only severity scores features (4 features)	4.29	2.85	19%	0.440	4.32	0.02	2.86	0.01	20%	0.005
Dataset not considering severity scores features (86 features)	4.01	2.67	29%	0.544	3.97	0.04	2.61	0.03	32%	0.009
Feature Selection analysis										
Dataset after feature selection (26 features of Rfe Treebag)	4.03	2.68	29%	0.536	3.99	0.03	2.61	0.01	31%	0.008
Dataset after feature selection (27 features of Rfe Random Forest)	4.03	2.68	29%	0.537	3.99	0.03	2.61	0.01	32%	0.002
Dataset after feature selection without severity scores (28 features of Rfe Treebag)	**4.03**	**2.68**	**29%**	**0.537**	**3.99**	**0.03**	**2.62**	**0.01**	**31%**	**0.005**
Dataset after feature selection without severity scores and including UnitCode (23 features of Rfe Treebag)	4.01	2.60	32%	0.564	3.96	0.04	2.58	0.02	31%	0.006
Missing data analysis										
Dataset after feature selection without features with more than 15% of missing (26 features of Rfe Treebag)	4.05	2.67	28%	0.532	4.02	0.03	2.64	0.01	31%	0.004
Dataset after feature selection without features with more than 3% of missing (26 features of Rfe Treebag)	4.07	2.68	27%	0.523	4.11	0.03	2.70	0.01	27%	0.005

Table 4.10: Analysis of different datasets of Feature Selection

As can be seen from Table 4.10, compared to the baseline scenario (RMSE = 3.99; MAE = 2.65; R^2 = 30%), the scenario with the dataset considering only severity scores did not perform well (RMSE = 4.29; MAE = 2.85; R^2 = 19%). This result shows that adjusting models using just severity score features could imply severe bias, which was one great reason for using SLOSR instead of SRU (as explained in Section 3.4). Although these severity scores represent combinations of several other variables, they did not fully capture the effect of all original features presented in our dataset. Regarding the scenario excluding the severity scores, we can see that the performance indicators remain similar to the baseline scenario (RMSE = 4.01; MAE = 2.67; R^2 = 29%). This similar result may be explained by the fact that the dataset without the severity scores still has 86 features, a great number of covariates to explain the ICU LoS.

Regarding the feature selection analysis, first we compared the Treebag-RFE to the RF-RFE using the complete dataset (90 features). The Treebag-

RFE selected 26 features with best importance to the model, while the RF-RFE selected 27. We also tested a scenario excluding the severity scores from the original dataset and applying the Treebag-RFE, which selected 28 features with the best importance. We can note that when we ran the model considering only the best importance variables of each feature selection scenario, all three previous scenarios presented similar results (RMSE = 4.03; MAE = 2.68; R^2 = 29%). We ran a feature selection scenario including the ICU code (and excluding the severity scores) to check if the ICUs should be considered or not. The results were also very close to the scenario without this feature (RMSE = 4.01; MAE = 2.60; R^2 = 32%). However, this scenario with the ICU code would be a problem for benchmarking purposes since we are adjusting the prediction considering possible ICU efficiencies or lack of efficiency, including bias in the benchmarking model. Therefore, since the performance difference between the model with and without this feature is low, we preferred not to include this feature.

We also did a sensitivity analysis excluding imputed features with more than 15% of missing and other excluding variables with more than 3% of missing, and then applying the Treebag-RFE to select the best set of features. The model without these imputed features performed worse than the other models that included these features. So, we preferred to remain with these features.

Therefore, after analyzing several sets of features with the help of feature selection techniques, we chose to use the 28 features selected by the Treebag-RFE from the dataset without the severity score features. This set of features presented a result similar to the one obtained by the scenario with all features (28 selected features: RMSE = 4.03, MAE = 2.68, R^2 = 29%; all 90 features: RMSE = 3.99; MAE = 2.65; R^2 = 30%). The relative importance of each selected feature is presented in Table 4.11. We will use this dataset with 28 features to compare all regression models to predict ICU LoS, since it will improve the models' efficiency and avoid overfitting.

Features	Importance
LowestGlasgowComaScale1h	100.00%
Urea	70.97%
AdmissionMainDiagnosisName	68.22%
Age	53.56%
IsMechanicalVentilation	50.55%
LengthHospitalStayPriorUnitAdmission	41.10%
HighestLeukocyteCount1h	39.75%
HighestCreatinine1h	38.37%
BMI	37.99%
HighestHeartRate1h	32.01%
Bilirubin	22.45%
HighestTemperature1h	19.34%
HighestRespiratoryRate1h	16.66%
n_complication	14.13%
AdmissionSourceName	14.06%
AdmissionTypeName	12.67%
IsVasopressors	10.80%
IsNonInvasiveVentilation	9.10%
IsDementia	4.52%
IsCrfNo	3.59%
FrailPatientMFI	3.55%
Gender	2.93%
ChfNyha	2.11%
IsAngina	1.34%
IsAcuteAtrialFibrilation	0.87%
IsAlcoholism	0.85%
IsAcuteKidneyInjury	0.70%
IsAids	0.20%

Table 4.11: Importance of selected features

4.3 Predicting the Numeric ICU Length of Stay

After preprocessing the dataset and selecting the best set of features, we applied the proposed methodology to predict the numeric ICU LoS (see Section 3.2). So, we ran and compared nine different types of regression models. The comparison results are presented in Table 4.12.

Models	Testing set				Training set					
	RMSE	MAE	R^2	Cor	RMSE	(SD)	MAE	(SD)	R^2	(SD)
Random Forest	3.84	2.58	0.35	0.596	3.90	0.02	2.59	0.01	0.35	0.003
Boosting (GBM)	4.03	2.68	0.29	0.537	3.99	0.03	2.62	0.01	0.31	0.005
Linear Regression	4.16	2.77	0.24	0.491	4.19	0.02	2.77	0.01	0.25	0.003
kNN	4.17	2.67	0.24	0.492	4.22	0.03	2.69	0.01	0.24	0.003
Glm Negative Binomial	4.17	2.75	0.24	0.487	4.20	0.04	2.75	0.02	0.24	0.009
CART	4.21	2.80	0.22	0.473	4.22	0.03	2.80	0.02	0.23	0.008
SVR Radial	4.24	2.49	0.21	0.502	4.29	0.03	2.50	0.01	0.25	0.006
Bagging	4.32	2.91	0.18	0.428	4.33	0.04	2.91	0.01	0.19	0.010
SVR Linear	4.34	2.56	0.17	0.478	4.38	0.03	2.56	0.01	0.23	0.004

Table 4.12: Statistical comparison between Regression Models

From Table 4.12 we can note that the best result was obtained by the Random Forests (RF) model (RMSE = 3.84; MAE = 2.58; R^2 = 0.35). The second best model was the GBM (RMSE = 4.03; MAE = 2.68; R^2 = 0.29), which is also a tree-based model. The other models presented inferior results compared to the best two models.

To compare the calibration of the best models, we analyzed the curve of predicted versus observed ICU LoS in the testing sample, as illustrated in Figures 4.6 and 4.7. We can note that, for patients with observed LoS bigger than ten days, Random Forests model starts to underestimate the ICU LoS. From the perspective of calibration, GBM seems to have a better result for patients with a longer length of stay, presenting a confidence interval of the predicted curve closer to the observed curve.

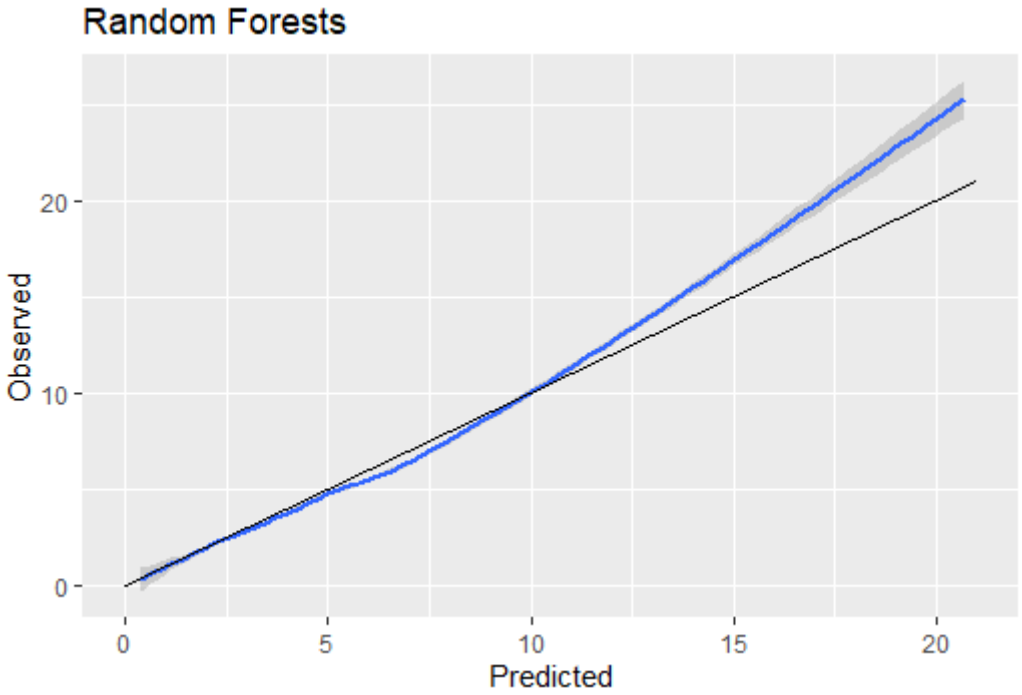


Figure 4.6: Calibration of Random Forests model. The predicted curve is in blue, and the perfect calibration curve is in back color.

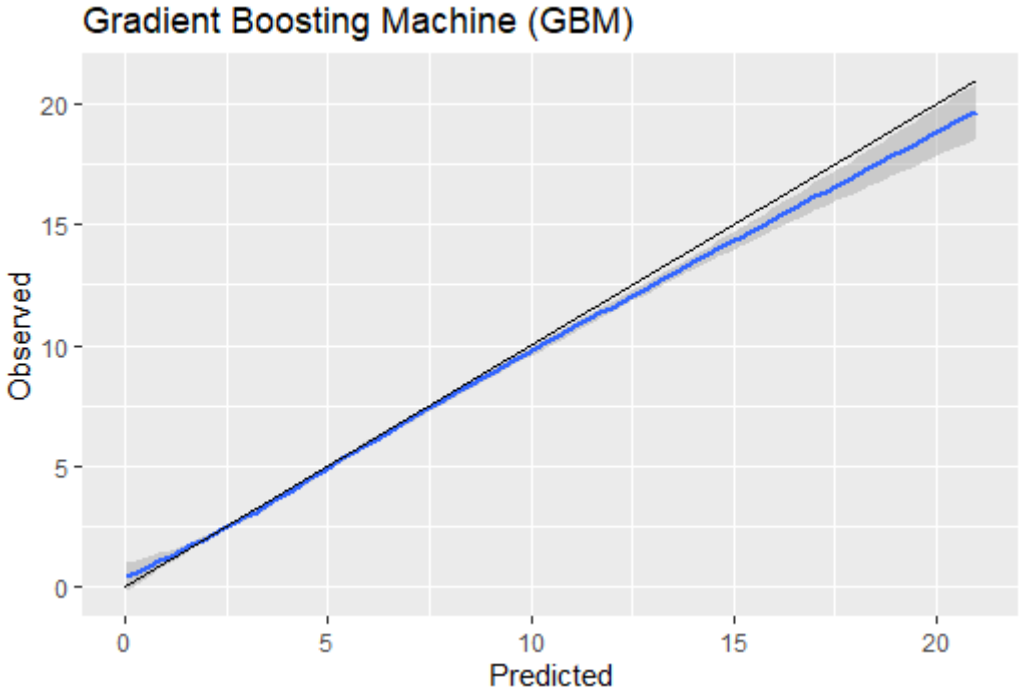


Figure 4.7: Calibration of GBM model. The predicted curve is in blue, and the perfect calibration curve is in back color.

Another relevant analysis to check the calibration of the models is the analysis of the prediction variability under ICU LoS stratum. We analyzed the distribution of ICU LoS predictions under the ranges of ICU LoS, as illustrated in Figures 4.8 and 4.9. We can see that both models presented a similar number of predicted cases (N Pred.) for all ranges of ICU LoS. Besides, both models overestimate the ICU LoS for the range "<=3" and underestimate for the ranges bigger than seven days (7-10; 10-14; >14). The predictions for the range "3-7" seems to be the best in both models. Moreover, the number of predictions for the range ">14" was very low compared to the observations in this range, which is related to the difficulty of predicting this type of patient.

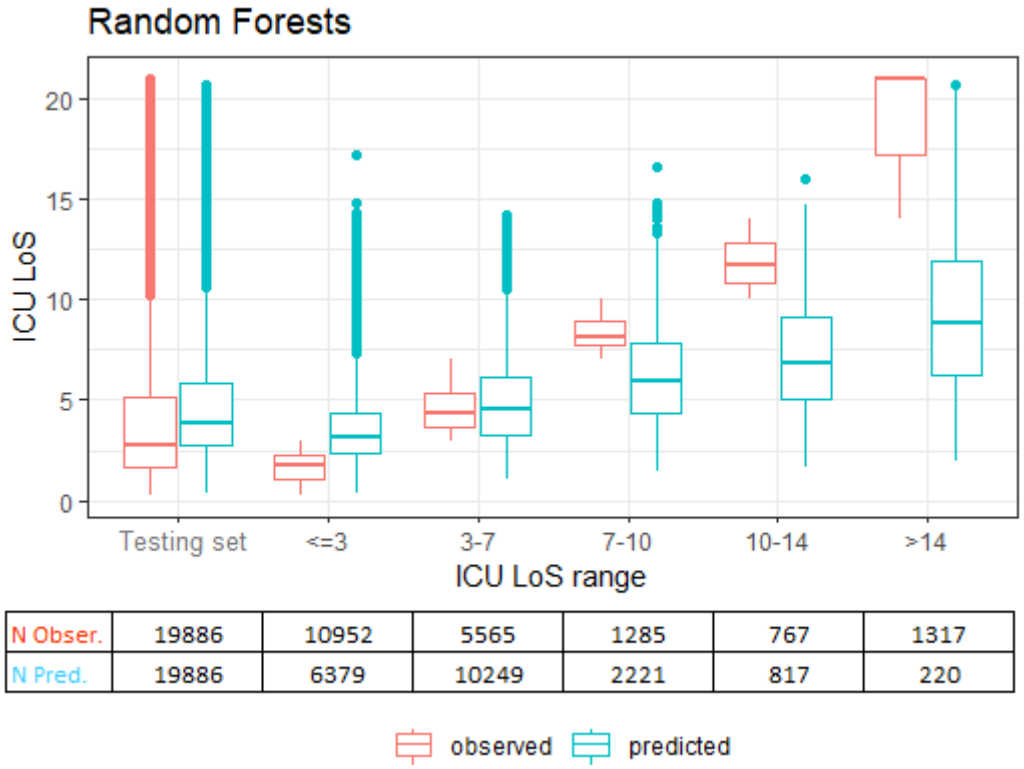


Figure 4.8: Boxplot of Random Forests model

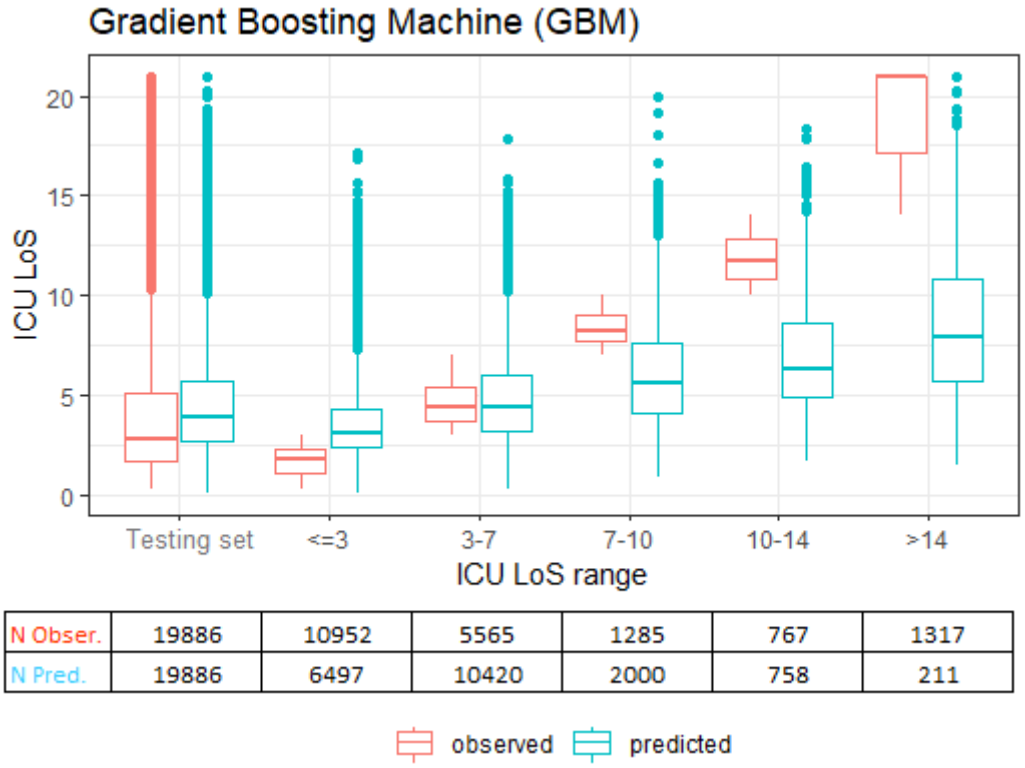


Figure 4.9: Boxplot of GBM model

Tables 4.13 and 4.14 presents the RMSE and the number of correctly predicted cases for each model. We can observe that RF shows a lower RMSE for all ICU LoS ranges compared to GBM, except for the range "<=3" in which the RMSE is almost the same for both models. Moreover, RF also presents a bigger proportion of correctly predicted cases for all LoS ranges. In terms of correctly predicted versus total observed cases (N Observ.), we can note that the best proportions of correctly predictions for RF model are in the ranges "<3" (46.7%) and "3-7" (64.5%), and the worst in ranges "10-14" (15.5%) and ">14" (14.7%). However, the proportion of correctly predicted cases versus total predicted cases was the biggest for this range ">14" (88.2%), which means that the model predicts few cases for this range, but the cases predicted have a high probability of being right.

Random Forests

ICU LoS Range	RMSE	Correctly Predicted / N Obser.	Correctly Predicted / N Pred.
All dataset	3.84	9364/19886 (47.1%)	9364/19886 (47.1%)
ICU LoS <=3	2.64	5116/10952 (46.7%)	5116/6379 (80.2%)
ICU LoS 3-7	2.25	3587/5565 (64.5%)	3587/10249 (35%)
ICU LoS 7-10	3.38	348/1285 (27.1%)	348/2221 (15.7%)
ICU LoS 10-14	5.46	119/767 (15.5%)	119/817 (14.6%)
ICU LoS >14	10.69	194/1317 (14.7%)	194/220 (88.2%)

Table 4.13: RF correct predictions for each ICU LoS range

GBM

ICU LoS Range	RMSE	Correctly Predicted / N Obser.	Correctly Predicted / N Pred.
All dataset	4.03	9223/19886 (46.4%)	9223/19886 (46.4%)
ICU LoS <=3	2.63	5155/10952 (47.1%)	5155/6497 (79.3%)
ICU LoS 3-7	2.37	3560/5565 (64%)	3560/10420 (34.2%)
ICU LoS 7-10	3.68	278/1285 (21.6%)	278/2000 (13.9%)
ICU LoS 10-14	5.80	94/767 (12.3%)	94/758 (12.4%)
ICU LoS >14	11.45	136/1317 (10.3%)	136/211 (64.5%)

Table 4.14: GBM correct predictions for each ICU LoS range

Therefore, from Figures 4.6 and 4.7, we noted that the calibration curve of GBM model seemed to be better compared to RF. However, when we analyzed the calibration by ranges of ICU LoS (Figures 4.8 and 4.9), we noted that both RF and GBM presented similar calibrations. The models showed almost the same proportion of predicted patients in each range and similar distribution of predicted ICU LoS, with an underestimation trend for ICU stays over seven days. Finally, when we analyzed the RMSE and the number of correct predictions in each range of ICU LoS (Tables 4.13 and 4.14), we observed that the RMSE of RF model was lower for all LoS ranges compared to GBM, and the proportion of correctly predicted cases was higher for all ranges. In short, the Random Forests model presented a lower general RMSE (3.84) compared to GBM (4.03), a lower RMSE for all ICU LoS ranges, and more correctly predicted patients for all ranges. In this way, we selected the Random Forests to be our prediction model.

4.4 Predicting the Risk of Prolonged ICU Stay

Another relevant information used by ICU managers is the risk of a patient being a prolonged stay. For that reason, we used the model with the best performance in our tests (Random Forests) to build two propositions of prediction model, as described in Section 3.3. The first alternative predicts the risk of a patient to stay over 14 days (model A), and the second one predicts the risk of a patient to stay over the 90% percentile of ICU stay distribution for his diagnosis group (model B).

Regarding the second alternative, we performed statistical and clinical analyses in our dataset and noted that the distribution of ICU LoS presented a great variability according to the patient diagnosis group. Table 4.15 shows the behavior of ICU LoS between the twenty most representative admission main diagnosis groups. For example, patients admitted with community-acquired pneumonia presented a mean ICU LoS equal to 6.7 days and a 90% percentile equal to 17.4 days, while patients with chest pain stayed on average 2.6 days and had a 90% percentile equal to 4.7 days. The complete table with the behavior of ICU LoS for all admission main diagnosis groups is presented in Table A.2. Therefore, since the expected length of stay may depend on the admission diagnosis, model B can be a good strategy to identify patients with critical conditions inside their diagnosis group and drive ICU care's immediate improvement.

Admission Main Diagnosis	Mean	P90	# of admissions
Community-acquired pneumonia	6.7	17.4	6988
Chest pain	2.6	4.7	4106
Symptomatic urinary tract infection, unspecified	5.7	13.3	3348
Syncope	3.4	6.0	3001
Unstable angina	3.1	5.8	2592
Acute (decompensated) heart failure	6.6	16.0	2079
Ischemic stroke	5.8	15.8	1666
Epilepsy and seizure disorders	4.5	10.0	1559
Atrial fibrillation	3.6	7.7	1346
Myocardial infarction without ST elevation	4.6	9.8	1204
Pulmonary thromboembolism	4.1	8.0	1144
Exogenous intoxications	2.6	4.7	1011
Transient ischemic accident	3.1	5.3	988
Gastroplasty	1.3	2.0	986
Gastroenteritis / gastroenterocolitis	3.6	7.5	935
Other diagnoses, not classified	3.8	7.9	923
High digestive bleeding	4.6	10.2	860
Traumatic brain injury	3.8	7.8	857
Other neurological complications	3.7	8.0	855
Decompensated COPD	6.6	15.0	844

Table 4.15: Behavior of ICU LoS between the most representative admission main diagnosis groups

Regarding the hyper-parameters, we used the best combination of parameters obtained in our tests (mtry = 6; minimal node size = 5) and tested three types of splitting rules (gini; extratrees; hellinger). The best training result was obtained by hellinger's splitting rule for both models, and the complete training results are presented in Tables B.1 and B.2. We trained the model with Random Forests and the results of the prediction for the testing set was as follows: (i) for model A, Brier Score = 0.05, AUC = 0.88, PPV = 0.87, NPV = 0.95; for model B, Brier Score = 0.08, AUC = 0.78, PPV = 0.86, NPV = 0.90. Other studies developed models to predict the risk of prolonged stay and reported the results in terms of AUC: Azari et al. [2012] reported 0.813, and Houthooft et al. [2015] reported 0.82, which were lower than our reported AUC for model A (0.88). Regarding the Brier Score, this metric simultaneously addresses calibration, consistency and sharpness, being a measure that varies in the range (0,1). Brier Score values close to zero represent more accurate models [Gneiting and Raftery, 2007; Rufibach, 2010]. Therefore, analyzing the Brier Score of both models, which were close to zero (model A: 0.05; model B: 0.08), and considering the results of the other measures (e.g., AUC = 0.88),

we can conclude that both models presented accurate predictions.

Figure 4.10 shows the confusion matrix for both models (considering a cutoff equal to 0.5). We can see that model B predicts more prolonged stay patients (368) compared to model A (327). However, the proportion of true positives was similar in both models (86%). We can also note that the models presented a different number of observed prolonged stay patients (model A: 1318; model B: 2192) because of the difference in definitions of prolonged stay. Since model B presented more patients defined as prolonged stay (which tend to be more difficult to predict), this may be one reason for worse outcomes than model A.

Confusion Matrix for Model A

	Reference	
Prediction	< 14 days	> 14 days
< 14 days	18524	1035
> 14 days	44	283

Confusion Matrix for Model B

	Reference	
Prediction	< P90	> P90
< P90	17643	1875
> P90	51	317

Figure 4.10: Confusion Matrix for Models A and B

We analyzed the calibration of each model by calibration belts, as illustrated in Figures 4.11 and 4.12. From Figure 4.11, we can see that, considering a confidence interval of 95%, model A predicts accurately until the probability of prolonged stay equal to 44%. After that, the model underestimates the risk of prolonged stay (prediction curve becomes above the red line). For model B, Figure 4.12 shows a similar behavior. However, the underestimation starts earlier, at the probability of prolonged stay equal to 31%, which indicates that model B underestimates the risk of prolonged stay for a higher range of predicted risk (31% - 100%) than model A (44% - 98%). In short, the predictions of risk in these ranges tend to be lower than the true risk of prolonged stay and should be analyzed with parsimonious.

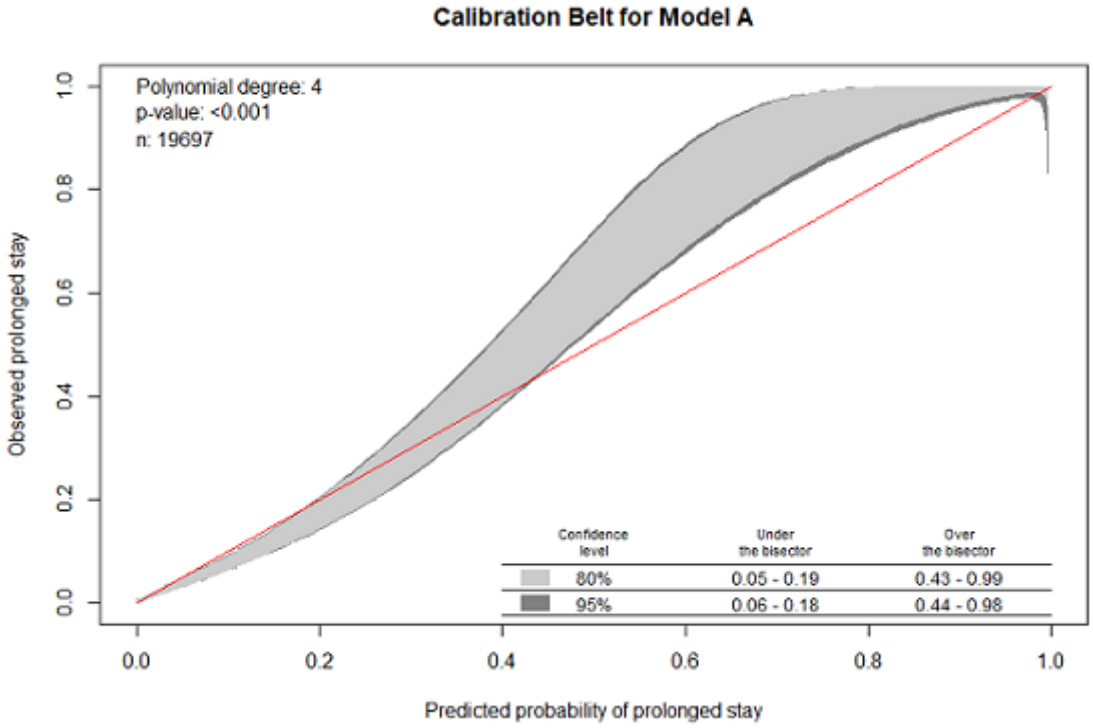


Figure 4.11: Calibration Belt for Model A

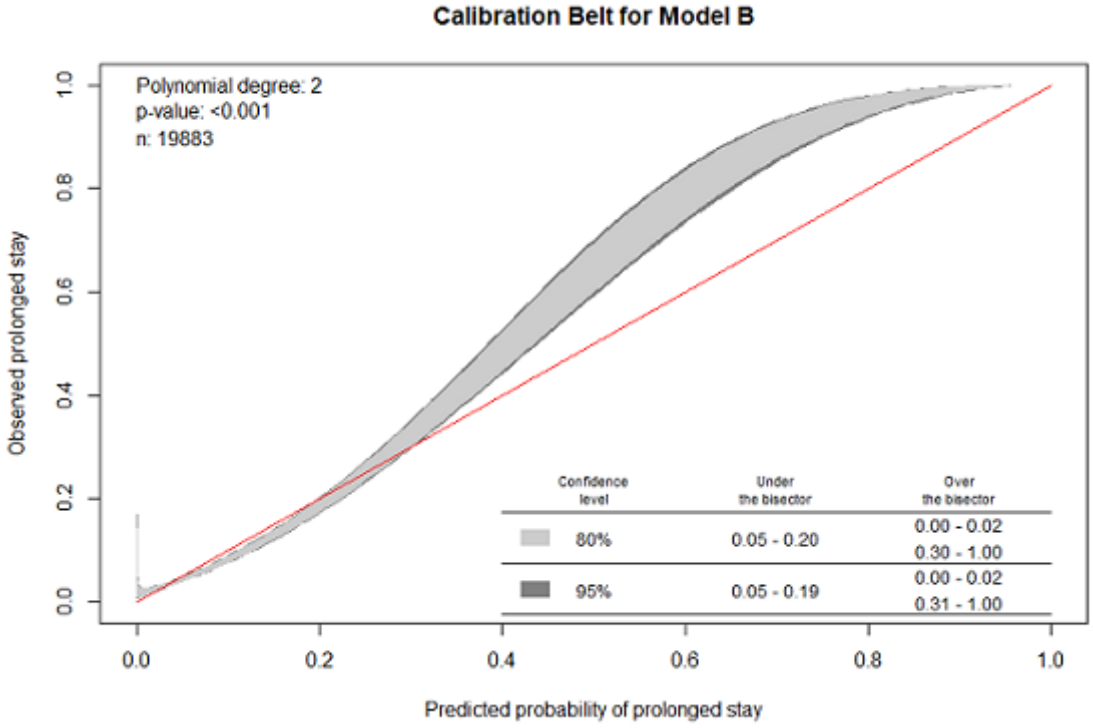


Figure 4.12: Calibration Belt for Model B

We can also analyze the number of observed high ICU LoS patients in each range of predicted risk for each model, as illustrated in Tables 4.16 and 4.17. The first, second, and third columns represent, respectively: the number of observed high ICU LoS patients included in each range of predicted risk, the number of predicted patients in each risk range, and the proportion of observed above the predicted ones. An accurate model should have as many high ICU LoS patients as possible inside the high-risk predicted ranges and minimize the number of high LoS patients inside the low-risk ranges. Analyzing the results of Tables 4.16 and 4.17, we can note that both models A and B presented similar behavior. The models accurately predict the high-risk patients since 97% of the predicted high-risk patients were really high ICU LoS patients. Moreover, the proportions of high ICU LoS patients inside the low and moderate risk ranges were small (2.1% and 15.4% for model A; 4.6% and 15.9%, for model B). Therefore, the analyses of both models show accurate results for the problem of predicting the risk of a patient to be a prolonged ICU LoS, and the decision about using each model depends on the main objective: to plan resources (model A) or to identify patients with critical conditions inside their diagnosis group (model B).

Range of predicted risk	Observed high ICU LoS patients	Number of predicted patients	%
Low [0-10%[	330	15782	2.1%
Moderate [10-33%[	513	3326	15.4%
High [33-67%[	341	640	53.3%
Very high [67-100%]	134	138	97.1%

Table 4.16: Model A - Number of observed high ICU LoS patients in each range of predicted risk

Range of predicted risk	Observed high ICU LoS patients	Number of predicted patients	%
Low [0-10%[	563	12195	4.6%
Moderate [10-33%[	1076	6784	15.9%
High [33-67%[	411	761	54.0%
Very high [67-100%]	142	146	97.3%

Table 4.17: Model B - Number of observed high ICU LoS patients in each range of predicted risk

4.5
Performing a Benchmarking Analysis between ICUs

We can also demonstrate how to use the proposed prediction model of Section 4.3 to perform a benchmarking analysis between ICUs. Figure 4.13 presents the comparison between the observed grouped length of stay per Unit and the predicted one. We can note an accurate calibration of the curve and a high coefficient of determination (R^2=0.93). Moreover, we can observe that the underestimation issue presented in the previous calibration curve of Figure 4.6 does not occur in the current analysis with grouped ICU LoS. The coefficient of determination reported for the individual prediction model of Section 4.3 was 0.35 compared to 0.93 for the grouped model, which also demonstrates the better calibration of the current model.

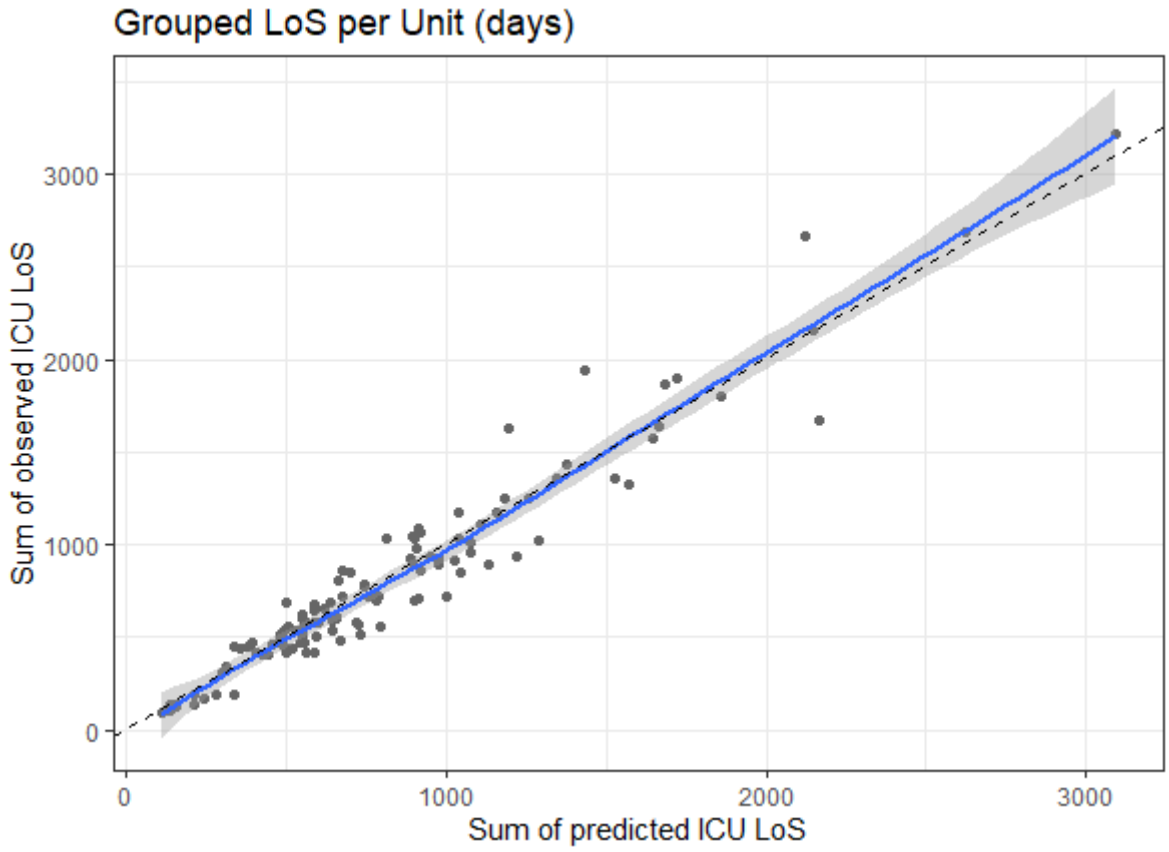


Figure 4.13: Calibration comparison between the observed grouped length of stay per Unit and the predicted one. The sum of observed ICU LoS is presented on the vertical axis and sum of predicted ICU LoS on the horizontal axis. Small dots represent ICUs and the solid line represents the reference value. Gray area represents confidence intervals.

Figure 4.14 presents a funnel plot for the Standardized Length of Stay Ratio (SLOSR) of each ICU. The overall SLOSR was equal to 0.99, which means an almost perfect calibration with a value very close to one. We could not compare our calibration to other papers because of the lack of measures reported in terms of R^2 and overall SLOSR. From Figure 4.14, we can see that 16 ICUs (14.7%) fall outside the 99.8% control limits and 33 ICUs (30.3%) outside the 95% control limits. So, 93 ICUs (85.3%) stayed inside the 99.8% control limits, performing according to the benchmark (considering the referred statistical confidence). These variability results reveal a proper calibration of the SLOSR model, in which a great proportion of ICUs stayed inside the control limits.

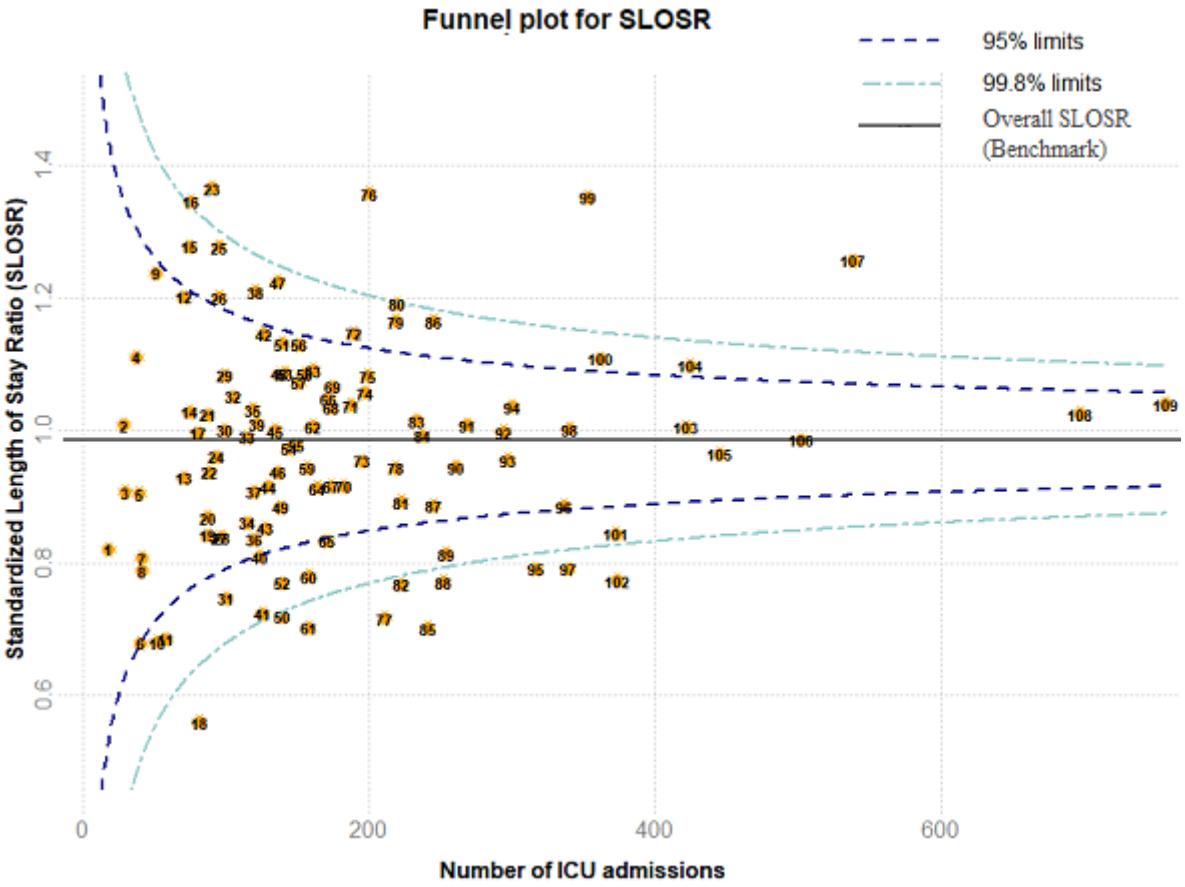


Figure 4.14: Funnel plot for the Standardized Length of Stay Ratio (SLOSR). The value of the quality indicator is presented on the vertical axis and the number of ICU admissions included when calculating the quality indicator is presented on the horizontal axis. Small dots with numbers represent ICUs and the solid line represents the benchmark value. Dashed lines represent control limits. Different types of dashed lines are used to differentiate between the 95% and 99.8% control limits.

We can also analyze the calibration of SLOSR for ICU types. So, we plotted the distribution of SLOSR for each ICU type (Figure 4.15). From this figure, we can see that the types of ICU with more admissions reported an almost perfect calibration, which is the case of General and Cardiac ICUs (median SLOSR=1, mean SLOSR=0.99). Surgical ICUs reported a lower median SLOSR (0.95), which shows a small trend of overestimation in our prediction model for ICU LoS. The dataset presented few admissions for Neurological, Oncological, and Orthopedic ICUs, which reported less than 400 cases each. For Orthopedic ICUs, our model presented an accurate calibration. However, the calibration for Neurological and Oncological types was not accurate. These types of ICUs will have a tendency to be efficient because of the calibration issue. This problem may be related to the small number of cases. Therefore, we do not recommend making any conclusions about the efficiency of the ICUs inside these two groups.

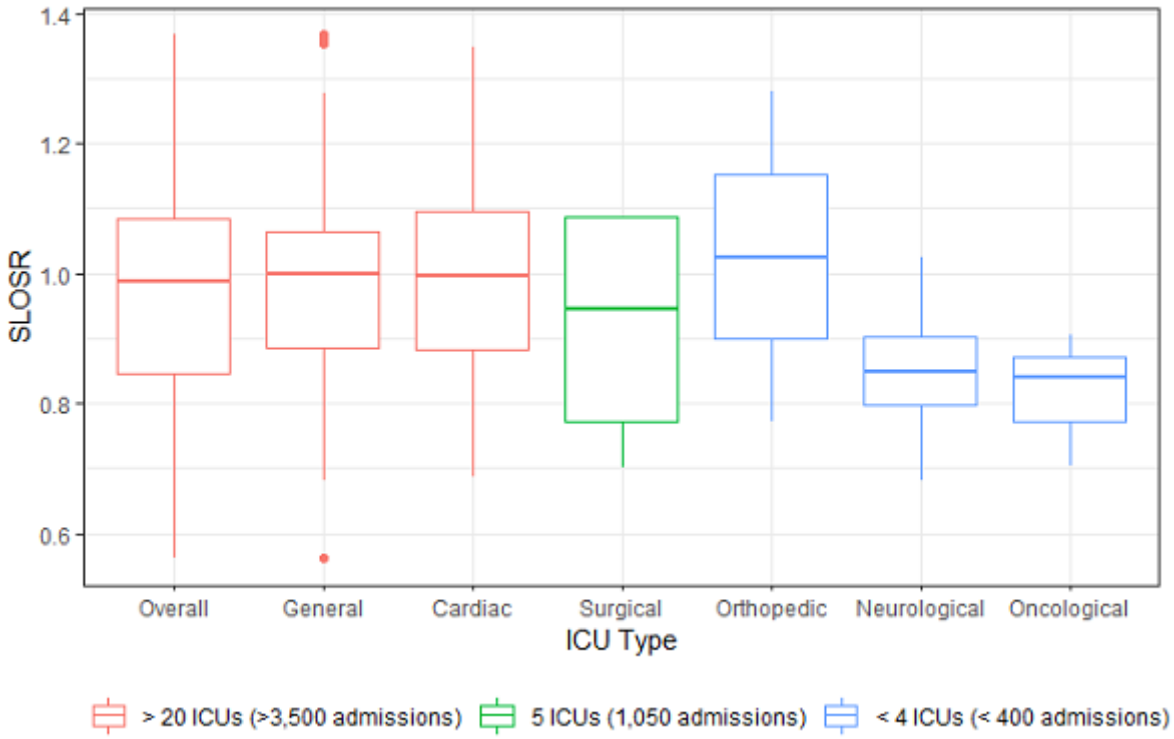


Figure 4.15: Boxplot for SLOSR for each ICU Type.

Regarding the efficiency analysis, using SLOSR measure we can find the efficient ICUs (SLOSR < 1) and the inefficient ones (SLOSR > 1). The most inefficient ICU was Unit 23 (SLOSR $= 1.37$), which means that the ICU used 37% more resources than expected. On the other side, we can see Unit 18 (SLOSR $= 0.56$), which can be noted as a very efficient ICU. The complete

table with SLOSR results for each ICU is presented in Supplementary Table C.1 and the general description of each ICU is presented in Supplementary Table C.2. Tables 4.18 and 4.19 show the main characteristics of the more efficient ICUs (SLOSR lower than 0.8) and more inefficient ones (SLOSR higher than 1.2), respectively. We can observe that the efficient ICUs from Table 4.18 showed the following characteristics: average age equal to 64.4 years old, average Glasgow equal to 14.3, and 6.6% of the patients were ventilated at admission. The admission type was 64.5% clinical, 28.8% elective surgery, and 6.7% urgent surgery. Regarding the inefficient ICUs from Table 4.19, the Units showed the following characteristics: average age equal to 66.3 years old, average Glasgow equal to 14.1, and 6.4% of the patients were ventilated at admission. The admission type was 84.7% clinical, 11.7% elective surgery, and 3.6% urgent surgery. We can note that inefficient ICUs presented, on average, older patients, with a lower Glasgow scale, and a small proportion of elective surgery patients compared to the efficient ICUs from Table 4.18, which seems to have a more severe case-mix. For this reason, the average expected ICU LoS predicted by our model was higher for this group compared to the efficient ones (5.7 versus 4.7) since our model considers the main features to predict an adjusted expected ICU LoS. So, despite presenting a more severe case-mix, the ICUs from Table 4.19 were considered inefficient because the observed ICU LoS was greater than the case-mix adjusted expected ICU LoS. Therefore, this information about efficient and inefficient ICUs can be used by managers to understand which ICU's organizational aspects could be associated with the ICU efficiency (e.g., the ratio of physicians per patient, number of beds, the process of care, etc.). Moreover, we can note that one Neurological (Unit 6) and one Oncological ICU (Unit 61) were considered inside the efficient group. However, we can not conclude anything about the efficiency of these ICUs because of the calibration issue shown in the previous analysis, which demonstrated that these types of ICUs have a tendency to be efficient.

UnitCode	ICU Type	SLOSR	# admissions	Average Age	Average Glasgow	% Mechanical Ventilation	% Clinical patients	% Elective Surgery patients	% Urgent Surgery patients	Average observed ICU LoS	Average expected ICU LoS
18	General	0.56	82	58.9	14.7	3.7%	67.1%	26.8%	6.1%	2.3	4.1
6	Neurological	0.68	40	69.4	13.1	10.0%	52.5%	30.0%	17.5%	3.6	5.3
10	General	0.68	52	61.5	13.8	11.5%	48.1%	50.0%	1.9%	3.7	5.4
11	Cardiac	0.69	58	55.2	14.9	0.0%	86.2%	12.1%	1.7%	2.9	4.2
85	Surgical	0.70	241	64.6	14.6	3.7%	16.2%	73.4%	10.4%	2.3	3.3
61	Oncological	0.70	158	57.7	14.5	5.7%	72.8%	17.7%	9.5%	3.3	4.6
77	General	0.72	211	68.0	14.3	4.3%	62.6%	27.5%	10.0%	3.4	4.7
50	Cardiac	0.72	139	67.1	14.7	2.9%	70.5%	22.3%	7.2%	3.0	4.2
41	General	0.72	126	68.3	14.2	6.3%	67.5%	24.6%	7.9%	3.9	5.3
31	General	0.75	100	71.6	14.1	6.0%	89.0%	11.0%	0.0%	4.2	5.7
82	General	0.77	223	72.6	14.4	6.7%	80.7%	17.5%	1.8%	4.2	5.5
52	Orthopedic	0.77	139	73.1	13.7	16.5%	93.5%	2.9%	3.6%	5.1	6.6
88	Surgical	0.77	252	64.6	14.7	2.4%	57.1%	37.7%	5.2%	2.8	3.6
102	General	0.77	374	59.6	13.6	9.1%	82.6%	7.8%	9.6%	4.5	5.8
60	Cardiac	0.78	158	67.1	14.1	18.4%	63.3%	34.2%	2.5%	3.6	4.6
8	General	0.79	41	65.5	14.3	4.9%	7.3%	70.7%	22.0%	2.7	3.4
97	General	0.79	339	56.5	14.8	3.2%	63.1%	33.9%	2.9%	2.6	3.3
95	General	0.79	317	57.4	14.6	2.8%	81.4%	18.6%	0.0%	3.2	4.1
Average		**0.73**	**169**	**64.4**	**14.3**	**6.6%**	**64.5%**	**28.8%**	**6.7%**	**3.4**	**4.7**

Table 4.18: Main characteristics of ICUs with SLOSR lower than 0.8 (more efficient ICUs)

UnitCode	ICU Type	SLOSR	# admissions	Average Age	Average Glasgow	% Mechanical Ventilation	% Clinical patients	% Elective Surgery patients	% Urgent Surgery patients	Average observed ICU LoS	Average expected ICU LoS
26	General	1.20	95	60.7	14.5	3.2%	72.6%	21.1%	6.3%	4.8	4.0
12	General	1.20	71	71.5	13.9	4.2%	80.3%	15.5%	4.2%	6.7	5.6
38	General	1.21	121	73.6	14.5	3.3%	78.5%	14.9%	6.6%	6.6	5.5
47	General	1.22	136	61.8	14.3	4.4%	99.3%	0.7%	0.0%	6.3	5.1
9	Cardiac	1.24	51	65.2	14.7	0.0%	92.2%	3.9%	3.9%	8.7	7.0
107	General	1.26	539	47.0	14.6	2.4%	94.4%	1.9%	3.7%	4.9	3.9
25	General	1.28	95	76.9	14.0	0.0%	96.8%	2.1%	1.1%	10.9	8.6
15	Orthopedic	1.28	75	77.9	12.0	30.7%	98.7%	1.3%	0.0%	11.5	9.0
16	Cardiac	1.35	76	66.1	14.4	11.8%	55.3%	40.8%	3.9%	6.0	4.5
99	General	1.35	353	57.7	14.6	1.4%	89.8%	7.6%	2.5%	5.5	4.1
76	General	1.36	200	67.6	13.6	11.0%	78.5%	13.5%	8.0%	8.1	6.0
23	General	1.37	90	70.1	14.5	4.4%	80.0%	16.7%	3.3%	7.6	5.6
Average		**1.28**	**159**	**66.3**	**14.1**	**6.4%**	**84.7%**	**11.7%**	**3.6%**	**7.3**	**5.7**

Table 4.19: Main characteristics of ICUs with SLOSR higher than 1.2 (more inefficient ICUs)

Therefore, we can conclude that our prediction model for grouped length of stay presented a high explanation and can be used to perform case-mix adjustments for benchmarking analysis between ICUs. We build a SLOSR measure using our prediction model as a source of information, and the model presented an almost perfect calibration (overall SLOSR equal to 0.99). Our model's accurate results show the importance of using the proposed data-driven methodology instead of standard literature models.

5
Conclusions

The main demands for ICU managers regarding the prediction of ICU LoS are: (i) planning the number of beds and staff required to fulfill the need for ICU care; (ii) identifying patients with a high risk of prolonged ICU LoS to drive immediate quality improvement; and (iii) enabling case-mix correction when comparing the LoS between ICUs (benchmarking). This thesis developed a structured data-driven methodology to approach each of these three clinical demands (explained in Sections 3.2, 3.3, and 3.4) and applied this methodology to a dataset with 109 mixed-type ICUs from 38 different Brazilian hospitals (Sections 4.3, 4.4, and 4.5). First, we proposed a model to predict the individual ICU length of stay, which can be used to plan the number of beds and staff required. Second, we proposed a model to predict the risk of prolonged stay, which helps identifying prolonged stay patients to drive immediate quality improvement. Finally, we used our prediction model for ICU LoS to build a case-mix adjusted measure (SLOSR) capable of performing non-biased benchmarking analyses between ICUs.

Regarding the prediction of individual ICU LoS, we compared nine regression models and concluded that Random Forests presented the best results (RMSE = 3.84; MAE = 2.58; $R^2 = 0.35$). We noted that the RMSE tends to increase for patients with higher observed ICU LoS, and the predictions for patients with observed high ICU LoS (>7days) tend to be underestimated. Besides, the predictions in the range "3-7 days" have the best calibration, and we noted a small prediction overestimation for short stays (< 3 days). Moreover, the number of predictions greater than 14 days was lower than the observations in this range, which is related to the difficulty of predicting this type of patient. However, the proportion of correctly predicted cases versus total predicted cases was high for these patients (88.2%), which means that the model predicts few cases for this range, but the predicted cases have a high probability of being right.

Regarding the prediction of the prolonged stay risk, we built two different models. One predicts the patient's risk of staying over 14 days (model A). The other predicts the risk of a patient staying over the 90% percentile of ICU stay distribution for his diagnosis group (model B). Model A follows the

type of classification defined by the literature. This model aims to identify the prolonged stay patients regardless of their diagnosis at admission, which is an important tool to plan the resources and operations required to fulfill ICU care. The results obtained by model A (Brier Score = 0.05, AUC = 0.88, PPV = 0.87, NPV = 0.95) reveals great accuracy compared to other studies (Azari et al. [2012] reported AUC = 0.813; and Houthooft et al. [2015] reported AUC = 0.82). We noted that it might not be reasonable to use just one threshold to define prolonged stay for all patients since the expected length of stay may depend on the admission diagnosis. For this reason, we proposed model B, which predicts a patient's risk to stay over the 90% percentile of ICU stay for his diagnosis group. To the best of our knowledge, no paper noted this relevant behavior. This model gives to the ICU managers more information about patients with critical conditions inside their diagnosis group, helping to improve their process of care. Model B presented similar results compared to model A. Both models accurately predict the high-risk patients since 97% of the predicted high-risk patients were really high ICU LoS patients. Moreover, the proportions of high ICU LoS patients inside the low and moderate risk ranges were small (2.1% and 15.4% for model A; 4.6% and 15.9%, for model B). Therefore, the analyses of both models show accurate results for the problem of predicting the risk of a patient to be a prolonged ICU LoS, and the decision about using each model depends on the main objective: to plan resources (model A) or to identify patients with critical conditions inside their diagnosis group (model B).

We also used our prediction model to build a case-mix adjusted measure (SLOSR) capable of performing non-biased benchmarking analyses between ICUs. The traditional measure used by ICUs to evaluate ICU efficiency is the Standardized Resource Use (SRU) proposed by Rothen et al. [2007]. This measure analyzes the observed and expected ICU length of stay to estimate the average amount of resources used in a specific unit. SRU uses SAPS3 to estimate the expected ICU LoS, being simple to be implemented and used by several managers and researchers to evaluate ICUs' efficiency [Bastos et al., 2020; Soares et al., 2015; Vincent et al., 2012; Wortel et al., 2021]. However, there is a limitation regarding how accurate is a prediction for ICU LoS using just the SAPS3 feature. In our application, we noted that limitation since the scenarios using only severity scores to predict ICU LoS performed worse than the scenario with the selected important features. For that reason, we used another measure of ICU efficiency, named Standardized Length of Stay Ratio (SLOSR) [Verburg et al., 2018a], which has a similar formulation compared to SRU. However, instead of using the SAPS3 to predict the expected ICU LoS, we

used our prediction model considering several relevant features, which improves case-mix correction and is essential when comparing the LoS between ICUs (benchmarking) [Marik and Hedman, 2000; Rapoport et al., 2003; Verburg et al., 2018a, 2017]. The prediction model of ICU LoS (grouped by ICUs) presented a high coefficient of determination (R^2=0.93) and can be used to perform non-biased case-mix adjustments for benchmarking analysis between ICUs. The SLOSR measure, built using our prediction model as a source of information, presented an almost perfect calibration (overall SLOSR equal to 0.99). Our model's accurate results show the importance of using the proposed data-driven methodology to build SLOSR instead of standard SRU. Therefore, SLOSR can be used to compare the general ICUs efficiency and also to make specific subgroups analyses, such as comparing the ICUs efficiency for each patient diagnosis.

In terms of regression models used to predict ICU LoS, our recent literature review analyzed six prediction articles that applied and compared different regression models to predict ICU LoS [Peres et al., 2021]. The authors noted that Support Vector Regression overcame the other models in two studies. Gradient Boosting Machine and Random Forests also presented superior results compared to other data-driven models. In our analysis, the Random Forests overcame SVR and GBM. Moreover, GBM was the model that presented the closest result to Random Forests (RMSE = 4.03; MAE = 2.68; $R^2 = 0.29$). GBM has clear similarities to Random Forests, since both estimate the final prediction throughout an ensemble of decision trees. However, the way the ensembles are constructed differs substantially between each method. In Random Forests, all trees are created independently and contribute equally to the final model, while in GBM new trees are dependent on past ones and contribute unequally to the final model. However, both models have a random process of resampling the dataset and the feature space to estimate several decision trees that will be further averaged [Breiman, 2001; Friedman, 2002]. This random process may be contributing to the superior results presented by these two regression models in our analysis, which is in line with previous literature.

Regarding the accuracy of the prediction model for the numeric ICU LoS, we can compare our performance indicators to the ones reported in other prediction studies. Verburg et al. [2014] compared six regression models to predict the ICU LOS for a dataset of 32,667 ICU admissions, and the Linear Regression presented the best result (RMSE = 7.28; $R^2 = 0.15$). Moran and Solomon [2012] compared seven regression models to predict the ICU LOS for a dataset of 111,663 ICU admissions, and the best model was the Linear Mixed

Model (LMM) (RMSE = 4.50; R^2 = 0.22). Houthooft et al. [2015] compared different data-driven models to predict the ICU LOS for patients remaining in the ICU on day 5, and the best performing model was SVR (R^2 = 0.22). Li et al. [2019] created a predictive model using preprocessing techniques and used Least Absolute Shrinkage and Selection Operator (LASSO) as prediction model (R^2 = 0.35). As we did not have the reported results in terms of RMSE for all papers, we compared the results based on the coefficient of determination (R^2). Therefore, the results obtained by Random Forests model using our proposed methodology presented a greater coefficient (R^2 = 0.35) compared to Verburg et al. [2014], Moran and Solomon [2012] and Houthooft et al. [2015], and similar results compared to Li et al. [2019].

Although our results reported better accuracy compared to the literature, we can observe that this coefficient is not good enough in terms of predictive explanation. The low R^2 values reported by most literature prediction studies may be explained by the fact of being difficult to predict the ICU LoS at the time of admission. Several relevant interventions are done after the patient admission, like the use of invasive devices, which may change the patient prognosis and evolution, being determinant to the understanding of ICU length of stay distribution. Then, we are planning for future studies to update the prediction for each day considering temporal variables (e.g., if the patient was not using a device and then turned to use; if the patient was using a device and then took it off; and if the patient was using a device and then continues to use).

In Chapter 2, we performed a systematic review and meta-analysis of risk factors for ICU LoS and concluded that the following features were potential risk factors: severity scores (e.g., APACHE and SAPS), Glasgow scale, BMI, admission source, admission type, readmission, inability to access Glasgow score, mechanical ventilation, clinical conditions (hypomagnesemia, delirium, malnutrition, infectious diseases, cerebrovascular accident, trauma, and respiratory diagnoses), chronic health items (COPD and chronic cardiovascular disease), reasons for ICU admission (sepsis, intracerebral hemorrhage, myocardial infarction, pulmonary edema, and subarachnoid hemorrhage) and clinical information (levels of red blood cell, body temperature, MR-proANP, albumin-creatinine ratio, and PaO2:FiO2 ratio) [Peres et al., 2020]. Our data-driven analysis started with 90 features, and after the preprocessing steps, we selected the 28 with the highest importance to the model (see Table 4.11). The most important feature reported in our model was Glasgow scale, which most literature also reported as a significant variable. Other significant features of the literature that was also important in our model were: admission

main diagnosis, admission source, admission type, BMI, mechanical ventilation, leukocyte count, body temperature, and creatinine. Urea, prior hospital length of stay, heart rate, bilirubin, and respiratory rate also reported high importance in our study and should be included in future prediction studies.

This thesis has the following limitations. The prediction models were trained considering the year 2019 and not consider possible case-mixes presented in other years. We did not include COVID-19 patients in this analysis, which might present a specific behavior that should be counted for future studies. The application represents a dataset of a big network of Brazilian private hospitals, which implies not having a complete case-mix context that includes public patients. We used Gradient Boosting Machine to perform the preprocessing sensitivity analyses, not running all scenarios for all models because of the problem's combinatorial nature. We trained the prediction model for the risk of prolonged stay using the best model (Random Forests) and set of hyper-parameters obtained in the previous tests for ICU LoS prediction, not considering all combinations of possible models and parameters. We were unable to assess the presence of possible discharge policies on certain days of the week, which is noted to occur in some Brazilian ICUs.

This thesis presented a structured data-driven methodology to approach three ICU managers' mains demands: planning the number of resources required, identifying patients with prolonged stay, and enabling non-biased benchmarking analyses between ICUs. The aim was not to deliver a standard model that fits all ICU types but rather a data-driven guide to generating predictions adjusted to the specific environment analyzed. Several studies have shown the importance of building data-driven guidelines to develop and validate models for routine use in clinical practice and reinforced the limitations of using standard models to any problem [Johnson et al., 2016; Rajkomar et al., 2019; Shillan et al., 2019]. For instance, the best result presented by Random Forests in our application does not mean that this model will perform well on any dataset. The recommendation is to follow the proposed methodology, test and compare the main predictive models, and check which one fits best. Preprocessing the dataset is an important step that should always be analyzed, with particular attention to feature selection techniques. The preprocessing analyses showed that some features needed to be transformed and others were not necessary for our model, reinforcing the need to analyze the dataset before running the tests. For future studies, we want to perform an external validation in another dataset of patients to check if our models could be used for other sets of hospitals. Moreover, we aim to evaluate the adherence of prediction models after including COVID-19 patients. Another extension for

this study would be to analyze prediction models for bed occupancy considering the ICU LoS prediction as a source of information. We also want to propose a model to predict the total patient time in all ICUs, considering the multiple patient admissions. Regarding the SLOSR, we want to compare our model to others, such as SRU. Finally, we want to clarify that predicting the length of stay at the time of ICU admission is not easy, and future studies should try to incorporate temporal features to the problem trying to get the evolution process of the patient.

Bibliography

Aggarwal, C. C. (2017). *Outlier Analysis*. Springer International Publishing, 2 edition.

Al Tehewy, M., El Houssinie, M., El Ezz, N. A., Abdelkhalik, M., and El Damaty, S. (2010). Developing severity adjusted quality measures for intensive care units. *International Journal of Health Care Quality Assurance*, 23(3):277–286. Number: 3 Reporter: International Journal of Health Care Quality Assurance.

Aloe, A. M. (2014). An empirical investigation of partial effect sizes in meta-analysis of correlational data. *The Journal of general psychology*, 141(1):47–64. Number: 1 Reporter: The Journal of general psychology.

Aloe, A. M. and Thompson, C. G. (2013). The synthesis of partial effect sizes. *Journal of the Society for Social Work and Research*, 4(4):390–405. Number: 4 Reporter: Journal of the Society for Social Work and Research.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Arabi, Y., Venkatesh, S., Haddad, S., Al Shimemeri, A., and Al Malik, S. (2002). A prospective study of prolonged stay in the intensive care unit: predictors and impact on resource utilization. *International Journal for Quality in Health Care*, 14(5):403–410. Number: 5 Reporter: International Journal for Quality in Health Care.

Atashi, A., Verburg, I. W., Karim, H., Miri, M., Abu-Hanna, A., Jonge, E. d., Keizer, N. F. d., and Eslami, S. (2018). Models to predict length of stay in the Intensive Care Unit after coronary artery bypass grafting: a systematic review. *The Journal of Cardiovascular Surgery*, 59(3):471–482. Number: 3 Reporter: The Journal of Cardiovascular Surgery.

Awad, A., Bader–El–Den, M., and McNicholas, J. (2017). Patient length of stay and mortality prediction: A survey. *Health Services Management Research*, 30(2):105–120. Number: 2 Reporter: Health Services Management Research.

Azari, A., Janeja, V. P., and Mohseni, A. (2012). Healthcare data mining: predicting hospital length of stay (phlos). *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)*, 3(3):44–66.

Bastos, L. S., Hamacher, S., Zampieri, F. G., Cavalcanti, A. B., Salluh, J. I., and Bozza, F. A. (2020). Structure and process associated with the efficiency of intensive care units in low-resource settings: An analysis of the checklist-icu trial database. *Journal of Critical Care*, 59:118–123.

Begg, C. B. and Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, pages 1088–1101. Reporter: Biometrics.

Bellia, C., Agnello, L., Lo Sasso, B., Bivona, G., Raineri, M. S., Giarratano, A., and Ciaccio, M. (2019). Mid-regional pro-adrenomedullin predicts poor outcome in non-selected patients admitted to an intensive care unit. *Clinical Chemistry and Laboratory Medicine*, 57(4):549–555. Number: 4 Reporter: Clinical Chemistry and Laboratory Medicine.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological*, pages 211–252.

Brascher, J., Peres, W., and Padilha, P. (2020). Use of the modified "nutrition risk in the critically ill" score and its association with the death of critically ill patients. *Clinical nutrition ESPEN*, 35:162–166.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Breslow, M. J. and Badawi, O. (2012a). Severity Scoring in the Critically Ill: Part 2: Maximizing Value From Outcome Prediction Scoring Systems. *Chest*, 141(2):518–527. Number: 2 Reporter: Chest.

Breslow, M. J. and Badawi, O. (2012b). Severity Scoring in the Critically Ill: Part 1—Interpretation and Accuracy of Outcome Prediction Scoring Systems. *Chest*, 141(1):245–252. Number: 1 Reporter: Chest.

Caetano, N., Laureano, R. M. S., and Cortez, P. (2014). A Data-driven Approach to Predict Hospital Length of Stay - A Portuguese Case Study. In *ICEIS*.

Cander, B., Dundar, Z. D., Gul, M., and Girisgin, S. (2011). Prognostic value of serum zinc levels in critically ill patients. *Journal of Critical Care*, 26(1):42–46. Number: 1 Reporter: Journal of Critical Care.

Chant, C., Smith, O. M., Marshall, J. C., and Friedrich, J. O. (2011). Relationship of catheter-associated urinary tract infection to mortality and length of stay in critically ill patients: A systematic review and meta-analysis of observational studies. *Critical Care Medicine*, 39(5):1167–1173. Number: 5 Reporter: Critical Care Medicine.

Chen, M., Sun, R., and Hu, B. (2015). The influence of serum magnesium level on the prognosis of critically ill patients. *Zhonghua Wei Zhong Bing Ji Jiu Yi Xue*, 27(3):213–217. Number: 3 Reporter: Zhonghua Wei Zhong Bing Ji Jiu Yi Xue.

Chevret, S., Seaman, S., and Resche-Rigon, M. (2015). Multiple imputation: a mature approach to dealing with missing data. *Intensive care medicine*, 41(2):348–350.

Choi, M. and Lee, H. S. (2016). Critical Patient Severity Classification System predicts outcomes in intensive care unit patients. *Nursing in Critical Care*, 21(4):206–213. Number: 4 Reporter: Nursing in Critical Care.

Cohen, J. (1988). *Statistical power analysis for the social sciences*. Routledge.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

da Silva, T. K., Berbigier, M. C., Rubin, B. d. A., Moraes, R. B., Corrêa Souza, G., and Schweigert Perry, I. D. (2015). Phase angle as a prognostic marker in patients with critical illness. *Nutrition in Clinical Practice: Official Publication of the American Society for Parenteral and Enteral Nutrition*, 30(2):261–265. Number: 2 Reporter: Nutrition in Clinical Practice: Official Publication of the American Society for Parenteral and Enteral Nutrition.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157.

Dong, G. and Liu, H. (2018). *Feature engineering for machine learning and data analytics*. CRC Press.

Ely, E. W., Gautam, S., Margolin, R., Francis, J., May, L., Speroff, T., Truman, B., Dittus, R., Bernard, R., and Inouye, S. K. (2001). The impact of delirium in the intensive care unit on hospital length of stay. *Intensive Care Medicine*, 27(12):1892–1900. Number: 12 Reporter: Intensive Care Medicine.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science: The Official Journal of the Society for Prevention Research*, 8(3):206–213.

Graybill, F. A. (1976). *Theory and application of the linear model*, volume 183. Duxbury press North Scituate, MA.

Gupta, S., Bindra, G., Batada, V., and Gupta, A. (2016). To study serum magnesium level and its correlation with prognostic significance in critically ill patients. *JIACM*, 17(1):36–39. Number: 1 Reporter: JIACM.

Hachesu, P. R., Ahmadi, M., Alizadeh, S., and Sadoughi, F. (2013). Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthcare informatics research*, 19(2):121.

Halpern, N. A. and Pastores, S. M. (2015). Critical care medicine beds, use, occupancy and costs in the United States: a methodological review. *Critical care medicine*, 43(11):2452. Number: 11 Reporter: Critical care medicine.

Halpern, N. A., Pastores, S. M., and Greenstein, R. J. (2004). Critical care medicine in the United States 1985–2000: an analysis of bed numbers, use, and costs. *Critical care medicine*, 32(6):1254–1259. Number: 6 Reporter: Critical care medicine.

Hayden, J. A., van der Windt, D. A., Cartwright, J. L., Côté, P., and Bombardier, C. (2013). Assessing Bias in Studies of Prognostic Factors. *Annals of Internal Medicine*, 158(4):280. Number: 4 Reporter: Annals of Internal Medicine.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., and Altman, D. G. (2003a). Measuring inconsistency in meta-analyses. *BMJ (Clinical research ed.)*, 327(7414):557–560. Number: 7414 Reporter: BMJ (Clinical research ed.).

Higgins, T. L. (2007). Quantifying risk and benchmarking performance in the adult intensive care unit. *Journal of intensive care medicine*, 22(3):141–156. Number: 3 Reporter: Journal of intensive care medicine.

Higgins, T. L., McGee, W. T., Steingrub, J. S., Rapoport, J., Lemeshow, S., and Teres, D. (2003b). Early indicators of prolonged intensive care unit stay: impact of illness severity, physician staffing, and pre-intensive care unit length of stay. *Critical Care Medicine*, 31(1):45–51. Number: 1 Reporter: Critical Care Medicine.

Houthooft, R., Ruyssinck, J., van der Herten, J., Stijven, S., Couckuyt, I., Gadeyne, B., Ongenae, F., Colpaert, K., Decruyenaere, J., Dhaene, T., and De Turck, F. (2015). Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores. *Artificial Intelligence in Medicine*, 63(3):191–207. Number: 3 Reporter: Artificial Intelligence in Medicine.

Jiang, P., Lv, Q., Lai, T., and Xu, F. (2017). Does Hypomagnesemia Impact on the Outcome of Patients Admitted to the Intensive Care Unit? A Systematic Review and Meta-Analysis:. *SHOCK*, 47(3):288–295. Number: 3 Reporter: SHOCK.

Johnson, A. E., Ghassemi, M. M., Nemati, S., Niehaus, K. E., Clifton, D. A., and Clifford, G. D. (2016). Machine learning and decision support in critical care. *Proceedings of the IEEE*, 104(2):444–466.

Kahn, J. M., Rubenfeld, G. D., Rohrbach, J., and Fuchs, B. D. (2008). Cost savings attributable to reductions in intensive care unit length of stay for mechanically ventilated patients. *Medical care*, pages 1226–1233. Reporter: Medical care.

Keegan, M. T., Gajic, O., and Afessa, B. (2011). Severity of illness scoring systems in the intensive care unit. *Critical Care Medicine*, 39(1):163–169. Number: 1 Reporter: Critical Care Medicine.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, 53(11):3735–3745.

Kishi, Y., Iwasaki, Y., Takezawa, K., Kurosawa, H., and Endo, S. (1995). Delirium in critical care unit patients admitted through an emergency room. *General Hospital Psychiatry*, 17(5):371–379. Number: 5 Reporter: General Hospital Psychiatry.

Knaus, W. A., Wagner, D. P., Zimmerman, J. E., and Draper, E. A. (1993). Variations in mortality and length of stay in intensive care units. *Annals of Internal Medicine*, 118(10):753–761. Number: 10 Reporter: Annals of Internal Medicine.

Kramer, A. A. and Zimmerman, J. E. (2010). A predictive model for the early identification of patients at risk for a prolonged intensive care unit length of

stay. *BMC medical informatics and decision making*, 10:27. Reporter: BMC medical informatics and decision making.

Kuhn, M. (2009). The caret package. *Journal of Statistical Software*, 28(5).

Kuhn, M. and Johnson, K. (2013). Data Pre-processing. In Kuhn, M. and Johnson, K., editors, *Applied Predictive Modeling*, pages 27–59. Springer, New York, NY.

Kuhn, M. and Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. CRC Press.

Kumar, S., Honmode, A., Jain, S., and Bhagat, V. (2015). Does magnesium matter in patients of Medical Intensive Care Unit: A study in rural Central India. *Indian Journal of Critical Care Medicine : Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine*, 19(7):379–383. Number: 7 Reporter: Indian Journal of Critical Care Medicine : Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine.

Kwak, S. K. and Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean journal of anesthesiology*, 70(4):407.

Laupland, K. B., Kirkpatrick, A. W., Kortbeek, J. B., and Zuege, D. J. (2006). Long-term mortality outcome associated with prolonged admission to the ICU. *Chest*, 129(4):954–959. Number: 4 Reporter: Chest.

Li, C., Chen, L., Feng, J., Wu, D., Wang, Z., Liu, J., and Xu, W. (2019). Prediction of Length of Stay on the Intensive Care Unit Based on Least Absolute Shrinkage and Selection Operator. *IEEE Access*, 7:110710–110721. Conference Name: IEEE Access.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gotzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., and Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*, 339(jul21 1):b2700–b2700. Number: jul21 1 Reporter: BMJ.

Lim, W., Whitlock, R., Khera, V., Devereaux, P. J., Tkaczyk, A., Heels-Ansdell, D., Jacka, M., and Cook, D. (2010). Etiology of troponin elevation in critically ill patients. *Journal of Critical Care*, 25(2):322–328. Number: 2 Reporter: Journal of Critical Care.

Limaye, C. S., Londhey, V. A., Nadkart, M. Y., and Borges, N. E. (2011). Hypomagnesemia in critically ill medical patients. *The Journal of the Association*

of Physicians of India, 59:19–22. Reporter: The Journal of the Association of Physicians of India.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Liu, P., Lei, L., Yin, J., Zhang, W., Naijun, W., and El-Darzi, E. (2006). Healthcare Data Mining: Prediction Inpatient Length of Stay. In *2006 3rd International IEEE Conference Intelligent Systems*, pages 832–837. ISSN: 1941-1294.

Luo, D., Wan, X., Liu, J., and Tong, T. (2018). Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range. *Statistical methods in medical research*, 27(6):1785–1805. Number: 6 Reporter: Statistical methods in medical research.

Makrygiannis, S. S., Rizikou, D., Patsourakos, N. G., Lampakis, M., Margariti, A., Ampartzidou, O. S., Sakellaridis, K., Tselioti, P., Pipilis, A., and Prekates, A. A. (2018). New-onset atrial fibrillation and clinical outcome in non-cardiac intensive care unit patients. *Australian Critical Care: Official Journal of the Confederation of Australian Critical Care Nurses*, 31(5):274–277. Number: 5 Reporter: Australian Critical Care: Official Journal of the Confederation of Australian Critical Care Nurses.

Marik, P. E. and Hedman, L. (2000). What's in a day? Determining intensive care unit length of stay. *Critical Care Medicine*, 28(6):2090–2093. Number: 6 Reporter: Critical Care Medicine.

Mayer, E. K., Bottle, A., Rao, C., Darzi, A. W., and Athanasiou, T. (2009). Funnel plots and their emerging application in surgery. *Annals of surgery*, 249(3):376–383.

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, 6(7):6. Number: 7 Reporter: PLoS Medicine.

Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307.

Moran, J. L., Bristow, P., Solomon, P. J., George, C., Hart, G. K., and Australian and New Zealand Intensive Care Society Database Management Committee (ADMC) (2008). Mortality and length-of-stay outcomes, 1993-2003, in the binational Australian and New Zealand intensive care adult patient database. *Critical Care Medicine*, 36(1):46–61. Number: 1 Reporter: Critical Care Medicine.

Moran, J. L. and Solomon, P. J. (2012). A review of statistical estimators for risk-adjusted length of stay: analysis of the Australian and new Zealand Intensive Care Adult Patient Data-Base, 2008-2009. *BMC medical research methodology*, 12:68. Reporter: BMC medical research methodology.

Morello, L. G., Dalla-Costa, L. M., Fontana, R. M., Netto, A. C. S. d. O., Petterle, R. R., Conte, D., Pereira, L. A., Krieger, M. A., and Raboni, S. M. (2019). Assessment of clinical and epidemiological characteristics of patients with and without sepsis in intensive care units of a tertiary hospital. *Einstein (São Paulo)*, 17(2).

Morid, M. A., Sheng, O. R. L., Kawamoto, K., Ault, T., Dorius, J., and Abdelrahman, S. (2019). Healthcare cost prediction: Leveraging fine-grain temporal patterns. *Journal of biomedical informatics*, 91:103113.

Muscedere, J., Waters, B., Varambally, A., Bagshaw, S. M., Boyd, J. G., Maslove, D., Sibley, S., and Rockwood, K. (2017). The impact of frailty on intensive care unit outcomes: a systematic review and meta-analysis. *Intensive care medicine*, 43(8):1105–1122. Number: 8 Reporter: Intensive care medicine.

Myers, R. H. and Myers, R. H. (1990). *Classical and modern regression with applications*, volume 2. Duxbury press Belmont, CA.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Nicolas, F., Le Gall, J. R., Alperovitch, A., Loirat, P., and Villers, D. (1987). Influence of patients' age on survival, level of therapy and length of stay in intensive care units. *Intensive Care Medicine*, 13(1):9–13. Number: 1 Reporter: Intensive Care Medicine.

Niskanen, M., Reinikainen, M., and Pettilä, V. (2009). Case-mix-adjusted length of stay and mortality in 23 Finnish ICUs. *Intensive Care Medicine*, 35(6):1060–1067.

Ouimet, S., Riker, R., Bergeron, N., Bergeon, N., Cossette, M., Kavanagh, B., and Skrobik, Y. (2007). Subsyndromal delirium in the ICU: evidence for a disease spectrum. *Intensive Care Medicine*, 33(6):1007–1013. Number: 6 Reporter: Intensive Care Medicine.

Peres, I. T., Hamacher, S., Cyrino Oliveira, F. L., Bozza, F. A., and Salluh, J. I. (2021). Prediction of icu length of stay: a concise review. *Revista Brasileira de Terapia Intensiva*.

Peres, I. T., Hamacher, S., Oliveira, F. L. C., Thomé, A. M. T., and Bozza, F. A. (2020). What factors predict length of stay in the intensive care unit? Systematic review and meta-analysis. *Journal of Critical Care*.

Piva, S., Dora, G., Minelli, C., Michelini, M., Turla, F., Mazza, S., D'Ottavi, P., Moreno-Duarte, I., Sottini, C., Eikermann, M., and Latronico, N. (2015). The Surgical Optimal Mobility Score predicts mortality and length of stay in an Italian population of medical, surgical, and neurologic intensive care unit patients. *Journal of Critical Care*, 30(6):1251–1257. Number: 6 Reporter: Journal of Critical Care.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358.

Rakow, T., Wright, R. J., Spiegelhalter, D. J., and Bull, C. (2015). The pros and cons of funnel plots as an aid to risk communication and patient decision making. *British Journal of Psychology*, 106(2):327–348.

Rapoport, J., Teres, D., Zhao, Y., and Lemeshow, S. (2003). Length of stay data as a guide to hospital economic performance for icu patients. *Medical care*, pages 386–397.

Reini, K., Fredrikson, M., and Oscarsson, A. (2012). The prognostic value of the Modified Early Warning Score in critically ill patients: a prospective, observational study. *European Journal of Anaesthesiology*, 29(3):152–157. Number: 3 Reporter: European Journal of Anaesthesiology.

Rothen, H. U., Stricker, K., Einfalt, J., Bauer, P., Metnitz, P. G., Moreno, R. P., and Takala, J. (2007). Variability in outcome and resource use in intensive care units. *Intensive care medicine*, 33(8):1329–1336.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Rufibach, K. (2010). Use of brier score to assess binary predictions. *Journal of clinical epidemiology*, 63(8):938–939.

Salluh, J. I., Soares, M., and Keegan, M. T. (2017). Understanding intensive care unit benchmarking. *Intensive care medicine*, 43(11):1703–1707.

Salluh, J. I., Soares, M., Teles, J. M., Ceraso, D., Raimondi, N., Nava, V. S., Blasquez, P., Ugarte, S., Ibanez-Guzman, C., Centeno, J. V., Laca, M., Grecco, G., Jimenez, E., Árias Rivera, S., Duenas, C., Rocha, M. G., and Delirium Epidemiology in Critical Care Study Group (2010). Delirium epidemiology in critical care (DECCA): an international study. *Critical Care (London, England)*, 14(6):R210. Number: 6 Reporter: Critical Care (London, England).

Schoffelen, A., Hofhuis, J., Rommes, J., and Spronk, P. (2008). Consumption of ICU resources by long-stay patients does not change over time: 10-year observation in a teaching hospital in The Netherlands. *Critical Care*, 12(Suppl 2):P525. Number: Suppl 2 Reporter: Critical Care.

Schoffelen, A. F., Hofhuis, J. G. M., Posthouwer, D., Rommes, J. H., and Spronk, P. E. (2010). Consumption of ICU resources by long-stay patients does not change over time: 9-year observation in a teaching hospital in the Netherlands. *Neth J Crit Care*, 14(5):313. Number: 5 Reporter: Neth J Crit Care.

Seaton, S. E., Barker, L., Jenkins, D., Draper, E. S., Abrams, K. R., and Manktelow, B. N. (2016). What factors predict length of stay in a neonatal unit: a systematic review. *BMJ Open*, 6(10):e010466. Number: 10 Reporter: BMJ Open.

Shillan, D., Sterne, J. A., Champneys, A., and Gibbison, B. (2019). Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. *Critical Care*, 23(1):1–11.

Soares, M., Bozza, F. A., Angus, D. C., Japiassú, A. M., Viana, W. N., Costa, R., Brauer, L., Mazza, B. F., Corrêa, T. D., Nunes, A. L., et al. (2015). Organizational characteristics, outcomes, and resource use in 78 brazilian intensive care units: the orchestra study. *Intensive care medicine*, 41(12):2149–2160.

Soliman, H. M., Mercan, D., Lobo, S. S. M., Mélot, C., and Vincent, J.-L. (2003). Development of ionized hypomagnesemia is associated with higher mortality rates. *Critical Care Medicine*, 31(4):1082–1087. Number: 4 Reporter: Critical Care Medicine.

Spiegelhalter, D. J. (2005). Funnel plots for comparing institutional performance. *Statistics in medicine*, 24(8):1185–1202.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ (Clinical research ed.)*, 338:b2393.

Straney, L. D., Udy, A. A., Burrell, A., Bergmeir, C., Huckson, S., Cooper, D. J., and Pilcher, D. V. (2017). Modelling risk-adjusted variation in length of stay among Australian and New Zealand ICUs. *PloS One*, 12(5):e0176570. Number: 5 Reporter: PloS One.

Thompson, S. G. and Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine*, 18(20):2693–2708. Number: 20 Reporter: Statistics in Medicine.

Thomé, A. M. T., Scavarda, L. F., and Scavarda, A. J. (2016). Conducting systematic literature review in operations management. *Production Planning & Control*, 27(5):408–420. Number: 5 Reporter: Production Planning & Control.

Thorevska, N., Sabahi, R., Upadya, A., Manthous, C., and Amoateng-Adjepong, Y. (2003). Microalbuminuria in critically ill medical patients: prevalence, predictors, and prognostic significance. *Critical Care Medicine*, 31(4):1075–1081. Number: 4 Reporter: Critical Care Medicine.

Tsuruta, R., Nakahara, T., Miyauchi, T., Kutsuna, S., Ogino, Y., Yamamoto, T., Kaneko, T., Kawamura, Y., Kasaoka, S., and Maekawa, T. (2010). Prevalence and associated factors for delirium in critically ill patients at a Japanese intensive care unit. *General Hospital Psychiatry*, 32(6):607–611. Number: 6 Reporter: General Hospital Psychiatry.

Vasilevskis, E. E., Kuzniewicz, M. W., Cason, B. A., Lane, R. K., Dean, M. L., Clay, T., Rennie, D. J., Vittinghoff, E., and Dudley, R. A. (2009). Mortality Probability Model III and Simplified Acute Physiology Score II. *Chest*, 136(1):89–101. Number: 1 Reporter: Chest.

Verburg, I. W., de Jonge, E., Peek, N., and de Keizer, N. F. (2018a). The association between outcome-based quality indicators for intensive care units. *PloS one*, 13(6):e0198522.

Verburg, I. W., Holman, R., Peek, N., Abu-Hanna, A., and de Keizer, N. F. (2018b). Guidelines on constructing funnel plots for quality indicators: a case study on mortality in intensive care unit patients. *Statistical methods in medical research*, 27(11):3350–3366.

Verburg, I. W. M., Atashi, A., Eslami, S., Holman, R., Abu-Hanna, A., de Jonge, E., Peek, N., and de Keizer, N. F. (2017). Which Models Can I Use to Predict Adult ICU Length of Stay? A Systematic Review*:. *Critical Care Medicine*, 45(2):e222–e231. Number: 2 Reporter: Critical Care Medicine.

Verburg, I. W. M., de Keizer, N. F., de Jonge, E., and Peek, N. (2014). Comparison of Regression Methods for Modeling Intensive Care Length of Stay. *PLoS ONE*, 9(10):e109684. Number: 10 Reporter: PLoS ONE.

Verburg, I. W. M., Holman, R., Dongelmans, D., de Jonge, E., and de Keizer, N. F. (2018c). Is patient length of stay associated with intensive care unit characteristics? *Journal of Critical Care*, 43:114–121. Reporter: Journal of Critical Care.

Vincent, J.-L., Takala, J., and Flaatten, H. (2012). Impact of reimbursement schemes on quality of care: a european perspective.

Wan, X., Wang, W., Liu, J., and Tong, T. (2014). Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC medical research methodology*, 14(1):135. Number: 1 Reporter: BMC medical research methodology.

Weissman, G., Hubbard, R., Ungar, L., Harhay, M., Greene, C., Himes, B., and Halpern, S. (2018). Inclusion of Unstructured Clinical Text Improves Early Prediction of Death or Prolonged ICU Stay*. *Critical Care Medicine*, 46(7):1125–1132. Number: 7 Reporter: Critical Care Medicine.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399.

Wortel, S. A., de Keizer, N. F., Abu-Hanna, A., Dongelmans, D. A., and Bakhshi-Raiez, F. (2021). Number of intensivists per bed is associated with efficiency of dutch intensive care units. *Journal of Critical Care*, 62:223–229.

Zafar, M. S. H., Wani, J. I., Karim, R., Mir, M. M., and Koul, P. A. (2014). Significance of serum magnesium levels in critically ill-patients. *International Journal of Applied and Basic Medical Research*, 4(1):34–37. Number: 1 Reporter: International Journal of Applied and Basic Medical Research.

Zampieri, F. G., Ladeira, J. P., Park, M., Haib, D., Pastore, C. L., Santoro, C. M., and Colombari, F. (2014). Admission factors associated with prolonged (>14 days) intensive care unit stay. *Journal of Critical Care*, 29(1):60–65. Number: 1 Reporter: Journal of Critical Care.

Zhang, X., Xuan, W., Yin, P., Wang, L., Wu, X., and Wu, Q. (2015). Gastric tonometry guided therapy in critical care patients: a systematic review and meta-analysis. *Critical Care*, 19(1):22. Number: 1 Reporter: Critical Care.

Zhang, Z., Pan, L., and Ni, H. (2013). Impact of delirium on clinical outcome in critically ill patients: a meta-analysis. *General Hospital Psychiatry*, 35(2):105–111. Number: 2 Reporter: General Hospital Psychiatry.

Zimmerman, J. E., Kramer, A. A., McNair, D. S., Malila, F. M., and Shaffer, V. L. (2006). Intensive care unit length of stay: Benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. *Critical Care Medicine*, 34(10):2517–2529. Number: 10 Reporter: Critical Care Medicine.

Complete data dictionary

A
Supplementary Results for Data Preparation and Preprocessing

This appendix provides the supplementary results for data preparation and preprocessing.

Features	Min	Q1 - 3*IQR	Q3 + 3*IQR	Max	Inferior Limit Applied	Superior Limit Applied	Replaced by NA (Inf. Limit)	Replaced by NA (Sup. Limit)
Age	16	0	181	108			0	0
LengthHospitalStayPriorUnitAdmission	0	0	4	60			0	0
CharlsonComorbidityIndex	0	0	8	19			0	0
MFIpoints	0	0	8	10			0	0
MFIScore	0	0	0.72	0.91			0	0
Saps3Points	4	0	102	134			0	0
Saps3DeathProbabilityStandardEquation	0.01	0	66.15	98.47			0	0
SofaScore	0	0	8	20			0	0
n_complication	0	0	0	24			0	0
LowestSystolicBloodPressure1h	0	11	242	314	11		8	0
LowestDiastolicBloodPressure1h	0	1	141	197	1		8	0
LowestMeanArterialPressure1h	0	14.33	166	208.33	14.33		9	0
LowestGlasgowComaScale1h	3	15	15	15			0	0
LowestPlateletsCount1h	0	0	575	2000			0	0
HighestHeartRate1h	0	0	161	288			0	0
HighestRespiratoryRate1h	0	4	32	99	4		16	0
HighestTemperature1h	25	34.8	37.6	41.7	30		19	0
HighestLeukocyteCount1h	0	0	28.55	500			0	0
HighestCreatinine1h	0	0	2.64	40			0	0
Urea	0	0	139	990			0	0
BUN	0	0	64.94	462.62			0	0
BMI	3.7	3.3	50.2	726.6	10	100	7	145
PaO2FiO2	1	0	1377.55	2866.67		1377	0	30
PH	6	7.01	7.78	8	7.01	7.78	123	11
Bilirubin	0	0	1.98	80			0	0
Lactate	0	0	7.1	40			0	0

Table A.1: Description of Outliers

Admission Main Diagnosis	Mean	P90	# of admissions
Community-acquired pneumonia	6.7	17.4	6988
Chest pain	2.6	4.7	4106
Symptomatic urinary tract infection, unspecified	5.7	13.3	3348
Syncope	3.4	6.0	3001
Unstable angina	3.1	5.8	2592
Acute (decompensated) heart failure	6.6	16.0	2079
Ischemic stroke	5.8	15.8	1666
Epilepsy and seizure disorders	4.5	10.0	1559
Atrial fibrillation	3.6	7.7	1346
Myocardial infarction without ST elevation	4.6	9.8	1204
Pulmonary thromboembolism	4.1	8.0	1144
Exogenous intoxications	2.6	4.7	1011
Transient ischemic accident	3.1	5.3	988
Gastroplasty	1.3	2.0	986
Gastroenteritis / gastroenterocolitis	3.6	7.5	935
Other diagnoses, not classified	3.8	7.9	923
High digestive bleeding	4.6	10.2	860
Traumatic brain injury	3.8	7.8	857
Other neurological complications	3.7	8.0	855
Decompensated COPD	6.6	15.0	844
Nosocomial pneumonia	9.5	21.0	797
Hypertensive emergencies	3.4	6.2	695
Acute respiratory failure or discomfort	6.4	16.8	691
Deep vein thrombosis	3.5	7.4	671
Lower gastrointestinal bleeding	4.6	9.7	662
Acute pancreatitis	4.6	9.8	641
Cineangiocoronariography with stent implantatio	2.4	4.9	609
Digestive tract obstruction / subocclusion	6.1	16.0	592
Skin infection	6.7	18.8	584
Hyponatremia	4.6	9.2	582
Lumbosacral spine surgery, arthrodesis	2.0	3.6	574
Sepsis and septic shock	7.5	21.0	561
Non-surgical abdominal infection	5.6	12.8	511
Bradyarrhythmias and heart blocks	5.1	12.7	500
Infection of indeterminate focus	6.8	17.2	486
Polytrauma	4.3	9.8	481
High symptomatic urinary tract infection, pyelone	4.2	9.6	458
Tracheobronchitis / tracheitis	5.6	13.0	452
Electrophysiological study / ablation	1.4	2.1	442
Other vessels, stent angioplasty	2.3	4.1	442
Exploratory laparotomy	6.3	17.4	412
Stable angina	2.9	5.6	410
Acute lung edema	7.8	21.0	405
Knee surgery, arthroplasty	1.7	3.1	398
Retosigmoidectomy	4.0	10.5	393
Hip surgery, arthroplasty	2.3	4.3	389
Acute chronic renal failure	5.4	10.9	384

Admission Main Diagnosis	Mean	P90	# of admissions
Other respiratory diseases	5.4	12.0	365
Supraventricular tachycardia	3.0	5.5	355
Fall	4.8	9.8	354
Other cardiovascular complications	3.8	8.0	352
Other digestive complications	4.4	10.1	331
Systemic arterial hypertension	4.0	8.5	330
Fever of undetermined origin	4.6	9.7	329
Unspecified myocardial infarction with ST elevati	5.4	13.6	329
Intestinal diverticular disease and diverticulitis	4.4	10.3	323
Cholecystectomy	3.1	5.8	321
Cervical spine surgery, arthrodesis	1.8	2.8	307
Diabetic ketoacidosis	4.0	6.7	298
Delirium	5.9	15.4	297
Asthmatic crisis	3.6	7.1	292
Infection of soft tissues	7.5	21.0	287
Atrial flutter	3.8	8.2	274
Right cardiac catheterization	3.8	8.8	272
Radical prostatectomy	1.9	3.2	271
Pleural effusion	6.2	14.7	261
Other anemias	4.6	11.1	251
Neurosurgery for brain or intracranial tumor	4.1	9.2	246
Other encephalopathies	2.7	4.8	239
Lobectomies / segmentectomies	2.8	4.8	238
Femur surgery, osteosynthesis	4.0	8.5	237
Hypotension	5.3	12.9	236
Meningitis / meningoencephalitis	4.1	10.6	236
Depression and mood disorders	2.3	4.9	235
Hypokalemia (Hypokalemia)	3.3	6.1	228
Intentionally self-inflicted injuries / attempted s	3.2	5.8	224
Left cardiac catheterization without stenting	3.4	6.0	220
Nephrectomies	2.2	4.1	215
Subarachnoid hemorrhage	8.0	21.0	214
Cesarean delivery	2.7	4.6	214
Myocardial revascularization with CPB	5.1	8.0	212
Hyperkalaemia (Hyperkalemia)	3.7	7.7	211
Cardiac arrhythmias, unspecified	3.5	7.1	208
Acute renal failure	5.8	14.7	203
Aorta and iliac vessels, stent angioplasty	2.5	5.0	201
Diabetes mellitus with complications	4.8	9.4	201
Subdural hematoma / hygroma	5.5	12.7	199
Eclampsia / pre-eclampsia	2.9	5.2	198
Dengue	3.4	5.8	197
Febrile neutropenia	5.8	14.3	182
Intraparenchymal hemorrhage	8.7	21.0	174
Hepatic encephalopathy	6.2	14.3	171
Pacemaker implantation (permanent)	2.2	4.3	169
Clinical monitoring / observation	2.8	4.5	164

Admission Main Diagnosis	Mean	P90	# of admissions
Influenza	2.9	4.7	163
Dehydration	5.2	12.5	162
Arterial embolization	2.1	3.9	155
Aspiration pneumonitis	7.3	19.6	155
Sinusitis	3.2	6.0	155
Partial colectomy	5.5	13.4	154
Ventricular cardiac tachyarrhythmias	4.0	8.2	154
Cholecystitis	5.3	11.5	151
Musculoskeletal trauma	5.8	14.2	147
Abdominal hysterectomy	2.8	5.8	136
Other infectious complications	6.1	15.7	136
Hypoglycemia	4.3	10.1	131
Primary bloodstream infection associated with ve	7.0	14.6	131
Eat / torpor	8.1	21.0	129
Hyperglycemia	4.1	7.4	128
Pneumothorax	4.5	9.2	127
Arterial thrombosis / embolism	6.6	16.2	127
Appendectomy	3.8	7.5	126
Other unclassified surgeries	2.8	5.1	121
Rhabdomyolysis	3.4	6.4	121
Carotid artery, stent angioplasty	2.6	4.1	119
Vascular surgery of lower limbs	3.2	8.8	118
Other renal and urinary tract complications	4.5	11.9	115
Transurethral resections	2.2	3.4	115
Femur surgery, other	4.3	9.3	111
Surgeries of the jaw and oral cavity	1.6	2.8	110
Hemicolectomies	5.0	14.5	110
Low symptomatic urinary tract infection, cystitis	5.9	13.0	109
Pericarditis	2.6	4.3	109
Vessels of extremities, angioplasty with stent	2.8	7.2	109
Infectious diarrhea	5.5	12.4	107
Decompressive laminectomy	2.2	3.6	107
Trauma, others	5.2	11.4	102
Anxiety and stress disorders	2.2	4.0	100
Brain and intracranial tumors	5.0	11.0	97
Pseudomembranous colitis	4.1	10.1	95
Insertion / removal of double J catheter	2.6	4.5	95
Viral infections	3.7	7.1	94
Cholangitis	5.6	14.4	93
Chest trauma	3.8	7.9	86
Aortic aneurysm, without mention of dissection	6.2	21.0	85
Surgical site infection	6.8	15.9	85
Myocarditis	3.3	5.0	83
Pharyngitis / pharyngotonsillitis / tonsillitis	2.7	5.0	82
Preoperative monitoring	4.4	8.8	81
Other neurosurgeries	3.7	9.6	79
Pelvic trauma	5.3	10.5	78

Admission Main Diagnosis	Mean	P90	# of admissions
Knee surgery, other	2.6	5.0	77
Correction of inguinal hernia	3.0	6.7	76
Pleuroscopies	2.9	6.0	76
Anaphylaxis	1.9	3.7	74
Decompensated chronic renal failure	5.2	9.9	74
Acute appendicitis	3.1	6.3	73
Subdural hematoma drainage	7.6	21.0	72
Face trauma	2.9	5.0	72
Arteriography	3.9	10.0	71
Total colectomy	6.1	16.1	70
Monitoring of clinical treatments	4.3	7.9	70
Ventriculo-peritoneal, atrial or pleural shunt	2.5	3.9	69
Pericardial effusion	4.8	8.8	68
Aortic aneurysm, correction with stent	3.4	7.8	67
Hepatectomies	5.2	13.8	67
Other liver and bile duct complications	5.5	13.4	67
PCR, asystole / bradycardia	7.6	21.0	65
Thrombocytopenia and purpura	5.9	11.7	65
Aortic stenosis	7.5	21.0	64
Neurosurgery, cerebral aneurysm	4.0	8.7	63
Oral and maxillofacial surgery	1.2	1.8	62
Other oncological complications	6.8	13.7	62
Gastroduodenopancreatectomy (Whipple surgery	8.2	21.0	61
Hemorrhages (except digestive)	4.5	10.7	61
Other supraventricular cardiac arrhythmias	3.1	8.5	61
Abdominal / pelvic abscess	7.2	18.9	60
Enterectomies	5.1	11.4	59
Subtotal gastrectomy	4.0	7.8	59
Other prostate surgeries	2.1	3.4	59
Other surgeries for liver, biliary tract and pancrea	3.6	7.9	59
Upper limb surgery, other	2.3	3.7	58
Hepatical cirrhosis	7.4	17.1	57
Aortic dissection	7.0	17.2	57
Acute hepatitis	4.3	8.2	57
Drug toxicity (except exogenous intoxication and	4.8	13.2	57
Lumbosacral spine surgery, other	3.2	6.6	55
Osteosynthesis of upper limbs	1.8	3.5	55
Outras cirurgias da coluna	1.8	3.7	55
Outras complicações hematológicas	4.9	12.1	55
Vertigem e transtornos da função vestibular	2.4	3.9	55
Osteomielite	8.9	21.0	54
Outras doenças neuromusculares	4.3	10.0	54
Endocardite	9.5	21.0	53
Cirurgia do ombro, artroplastia	1.6	2.9	52
Queimadura corporal	5.8	15.9	52
Anemia falciforme e crise falcêmica	5.8	10.6	51
Outras cirurgias ginecológicas	3.1	4.3	51

Admission Main Diagnosis	Mean	P90	# of admissions
Cirurgias do quadril, outras	4.3	11.0	50
Outras cirurgia de cabeça e pescoço	3.4	8.0	50
Biópsia pulmonar (toracotomia ou VATS)	2.4	4.9	49
Pancreatectomias	4.0	7.8	49
Trombose de seio venoso	2.9	5.8	49
Encefalites	4.5	7.8	48
Hipocalcemia	4.0	7.7	48
Outras doenças psiquiátricas	4.3	8.8	48
Prostatite	5.3	21.0	48
Tromboembolectomia de vasos periféricos	4.3	8.7	48
Úlcera péptica / gastrite	2.5	4.8	48
Decorticação pulmonar	2.7	5.0	47
Implante de cateteres vasculares	3.8	8.4	47
Troca valvar aórtica	5.1	8.5	47
Cirurgia coluna torácica, artrodese	3.4	7.5	46
Hemoptise	5.7	17.4	46
Tuberculose pulmonar	4.9	11.9	46
Varicela / Herpes Zoster	5.8	12.3	46
Gastrectomia total	6.0	20.5	45
Peritonite	7.0	16.7	45
Síndrome de Guillain-Barré	5.6	10.9	45
Cirurgias do ureter	1.9	3.0	44
Doença litiásica renal	3.2	7.6	44
Neutropenia	7.1	19.2	44
Sarampo	3.4	5.7	44
Tireoidectomias	2.1	3.9	43
Outras doenças osteoarticulares	3.4	5.3	42
Pancitopenia	5.6	10.1	42
Laparoscopia	3.4	6.0	41
Linfadenectomias	2.5	3.1	41
Miocardiopatias	4.4	9.5	41
Ascite	6.2	10.9	40
Choque cardiogênico	10.8	21.0	40
Implante de cardiodesfibrilador	2.1	3.8	40
Osteossínteses de membros inferiores	3.8	10.7	40
Outras cirurgias vasculares	2.6	3.5	40
Correção de hérnias incisionais	4.2	9.0	39
Fechamento de colostomia	4.3	10.4	39
Lesão por inalação de fumaça	8.1	21.0	39
Pneumoperitôneo	6.3	15.9	39
Correção de comunicação interatrial (CIA)	2.4	4.2	38
Correções de outras hérnias	3.1	5.2	38
Endarterectomia de carótidas	3.1	5.2	38
Histerectomia por via vaginal	2.7	4.1	38
Cirurgias plásticas	1.2	2.2	37
PCR, não especificada	10.5	21.0	37
Septoplastia	1.4	2.8	37

Admission Main Diagnosis	Mean	P90	# of admissions
Vasos de extremidades, angioplastia sem stent	3.1	6.7	37
Artroscopia do ombro	1.4	2.9	36
Choque anafilático	2.6	4.4	36
Doença inflamatória intestinal (doença de Crohn	5.6	16.8	36
Hipercalcemia	5.4	5.9	36
Outros procedimentos invasivos cardíacos e vascu	2.4	5.1	36
Troca valvar mitral	5.7	12.4	36
Tumor de pâncreas	5.1	12.6	36
Insuficiência respiratória crônica	8.0	13.0	35
Isquemia enteromesentérica	4.9	8.6	35
Outras cirurgias / procedimentos endovasculares	2.5	3.8	35
Choque hipovolêmico	7.3	19.7	34
Confecção de fístula artério-venosa	3.1	6.6	34
Outras cirurgias ortopédicas	2.2	3.6	34
Polirradiculopatias / polineuropatias (exceto Guil	5.0	12.8	34
Outras doenças da pele	3.4	6.9	33
Quimioembolização arterial	2.5	3.8	33
Revisão de atrodese ou tratamento cirúrgico da p	2.1	4.3	33
Tumor de pulmão	6.6	18.6	33
Amputações de membros inferiores	3.2	6.9	32
Debridamento de tecidos moles e pele, outros	6.9	19.2	32
Diabetes mellitus sem complicações	4.3	8.5	32
Insuficiência respiratória por tumor	5.7	13.1	32
Outros vasos, angioplastia sem stent	2.5	7.4	32
Trombose de veia porta	3.7	6.0	32
Aneurisma cerebral	8.2	21.0	31
Farmacodermia	3.4	6.9	31
Outras cirurgias da cavidade abdominal	3.3	7.7	31
Transplante hepático	6.7	12.8	31
Acidose metabólica	6.5	14.9	30
Choque hemorrágico	5.5	16.2	30
Cirurgias da bexiga, outras	2.9	6.0	30
Insuficiência hepática aguda	5.7	11.5	30
Oclusão CIA / CIV / CA com prótese	1.9	3.0	30
Outras cirurgias das vias urinárias	2.9	4.8	30
Valvoplastia aórtica percutânea	4.1	6.4	30
Angiografia	4.0	11.8	29
Cirurgia de membros inferiores, outras	3.8	9.5	29
Esquizofrenia e transtornos psicóticos	3.7	6.9	29
Ferimento por projetil de arma de fogo	8.1	21.0	29
Hipertensão intracraniana	3.6	7.0	29
Infecção de corrente sanguínea primária não asso	10.7	21.0	29
Perfuração do trato digestivo	6.2	15.1	29
Ressecção de tumor retroperitoneal	4.6	10.7	29
Abstinência alcoólica / Delirium tremens	3.3	7.8	28
Anemias hemolíticas e hemólise	4.9	8.5	28
Colestase	5.4	10.2	28

Admission Main Diagnosis	Mean	P90	# of admissions
Drenagem de vias biliares	4.7	11.2	28
Esplenectomia ou esplenorrafia	3.6	6.4	28
Neurocirurgia de origem vascular, outras	4.9	14.5	28
Outras cirurgias gástricas	4.8	18.1	28
PCR, atividade elétrica sem pulso	10.9	21.0	28
Síndrome do baixo débito cardíaco	4.3	7.1	28
Artroscopia do joelho	1.5	3.3	27
Correção de hérnia umbilical	1.6	3.6	27
Demências	5.4	16.9	27
Endoscopia digestiva alta	5.5	12.2	27
Fibrose pulmonar	6.9	18.3	27
Insuficiência mitral	9.6	21.0	27
Neurocirurgias da medula espinhal	4.7	16.7	27
Ressecção de tumor de tecidos moles e pele	4.0	14.5	27
Suprarrenalectomia (adrenalectomia)	2.8	4.9	27
Transtornos devidos ao uso de substâncias psicoa	2.7	5.6	27
Tratamento cirúrgico da endometriose	2.2	3.5	27
Cirurgias da maxila	2.1	3.2	26
Complicações por tumores sólidos	6.1	20.8	26
Miastenia gravis	5.5	14.3	26
Ooforectomia	2.7	6.3	26
Outros distúrbios hidro-eletrolíticos	5.3	10.9	26
Troca de valva aórtica	8.4	19.7	26
Cirurgia coluna cervical, outras	1.5	2.0	25
Esclerose múltipla	3.8	7.5	25
Hidrocefalia	5.0	14.6	25
Trauma raquimedular	5.4	15.6	25
Drenagem de abscesso de tecidos moles e pele	5.2	18.1	24
Mastectomia	2.1	2.7	24
Outras doenças do aparelho genital feminino	2.5	3.8	24
Tumor de intestino grosso / canal anal	6.9	19.4	24
Artrite séptica	5.4	14.9	23
Cistectomias	5.8	12.8	23
Outras cirurgias do pulmão	2.6	4.1	23
Pancreatite crônica	5.4	12.2	23
Ressecção da massa mediastinal	1.9	3.0	23
Cirurgia aorta abdominal infrarrenal, aneurisma	4.2	8.2	22
Flebites e tromboflebites	5.1	13.5	22
Abscesso renal	4.0	7.9	21
Desnutrição grave e deficiências nutricionais	6.7	14.9	21
Endoscopia digestiva baixa / colonoscopia / retos	1.7	3.6	21
Hepatites crônicas	4.2	4.7	21
Síndrome edemigênica	4.2	9.0	21
Artroscopia do quadril	2.2	4.3	20

Table A.2: Behavior of ICU LoS for all admission main diagnosis groups

B
Supplementary Results for the Risk of Prolonged Stay Prediction

This appendix provides the supplementary results for the risk of prolonged stay prediction.

Run	Run 1	Run 2	Run3
mtry	6	6	6
min.node.size	5	5	5
splitrule	gini	extratrees	hellinger
BrierScore	0.0477	0.0485	0.0472
logLoss	0.1793	0.1788	0.1779
AUC	0.8649	0.8627	0.8686
prAUC	0.7196	0.7253	0.7213
Accuracy	0.9430	0.9402	0.9436
Kappa	0.3058	0.2581	0.3146
F1	0.3238	0.2757	0.3325
Sensitivity	0.1997	0.1668	0.2058
Specificity	0.9975	0.9968	0.9977
Pos_Pred_Value	0.8557	0.7954	0.8677
Neg_Pred_Value	0.9445	0.9423	0.9449
Precision	0.8557	0.7954	0.8677
Recall	0.1997	0.1668	0.2058
Detection_Rate	0.0136	0.0114	0.0141
Balanced_Accuracy	0.5986	0.5818	0.6017
logLossSD	0.0022	0.0011	0.0035
AUCSD	0.0042	0.0031	0.0030
prAUCSD	0.0066	0.0038	0.0052
AccuracySD	0.0009	0.0006	0.0010
KappaSD	0.0138	0.0110	0.0159
F1SD	0.0140	0.0114	0.0161
SensitivitySD	0.0104	0.0082	0.0120
SpecificitySD	0.0005	0.0005	0.0007
Pos_Pred_ValueSD	0.0269	0.0239	0.0332
Neg_Pred_ValueSD	0.0007	0.0005	0.0008
PrecisionSD	0.0269	0.0239	0.0332
RecallSD	0.0104	0.0082	0.0120
Detection_RateSD	0.0007	0.0006	0.0008
Balanced_AccuracySD	0.0052	0.0040	0.0060
BrierScoreSD	0.0005	0.0003	0.0005

Table B.1: Complete training results for risk prediction (model A)

Run	Run 1	Run 2	Run3
mtry	6	6	6
min.node.size	5	5	5
splitrule	gini	extratrees	hellinger
BrierScore	0.0817	0.0824	0.0811
logLoss	0.2909	0.2913	0.2880
AUC	0.7595	0.7591	0.7621
prAUC	0.6813	0.6743	0.6833
Accuracy	0.9034	0.9009	0.9037
Kappa	0.2057	0.1695	0.2078
F1	0.2285	0.1896	0.2306
Sensitivity	0.1318	0.1068	0.1329
Specificity	0.9975	0.9977	0.9976
Pos_Pred_Value	0.8641	0.8488	0.8725
Neg_Pred_Value	0.9041	0.9016	0.9042
Precision	0.8641	0.8488	0.8725
Recall	0.1318	0.1068	0.1329
Detection_Rate	0.0143	0.0116	0.0144
Balanced_Accuracy	0.5646	0.5522	0.5653
logLossSD	0.0033	0.0027	0.0029
AUCSD	0.0083	0.0070	0.0075
prAUCSD	0.0073	0.0072	0.0063
AccuracySD	0.0011	0.0012	0.0009
KappaSD	0.0175	0.0175	0.0142
F1SD	0.0190	0.0189	0.0155
SensitivitySD	0.0125	0.0118	0.0103
SpecificitySD	0.0005	0.0006	0.0005
Pos_Pred_ValueSD	0.0162	0.0298	0.0174
Neg_Pred_ValueSD	0.0012	0.0012	0.0010
PrecisionSD	0.0162	0.0298	0.0174
RecallSD	0.0125	0.0118	0.0103
Detection_RateSD	0.0014	0.0013	0.0011
Balanced_AccuracySD	0.0061	0.0058	0.0050
BrierScoreSD	0.0009	0.0008	0.0008

Table B.2: Complete training results for risk prediction (model B)

C
Supplementary Results for the Benchmarking Analysis between ICUs

This appendix provides the supplementary results for the Benchmarking Analysis between ICUs.

UnitCode	# of admissions	Sum of observed LoS	Sum of predicted LoS	SLOSR
1	18	92	111.98	0.82
2	29	141	139.82	1.01
3	30	142	156.56	0.91
4	37	347	312.15	1.11
5	39	193	213.35	0.90
6	40	143	210.44	0.68
7	41	127	157.25	0.81
8	41	110	139.46	0.79
9	51	444	358.57	1.24
10	52	191	280.72	0.68
11	58	169	246.17	0.69
12	71	474	394.14	1.20
13	71	411	442.23	0.93
14	75	308	299.39	1.03
15	75	861	673.12	1.28
16	76	456	338.68	1.35
17	81	473	474.29	1.00
18	82	190	339.07	0.56
19	87	440	521.92	0.84
20	87	476	547.55	0.87
21	87	418	407.85	1.02
22	88	515	549.12	0.94
23	90	686	501.97	1.37
24	93	409	425.75	0.96
25	95	1039	813.04	1.28
26	95	456	379.45	1.20

UnitCode	# of admissions	Sum of observed LoS	Sum of predicted LoS	SLOSR
27	95	437	520.55	0.84
28	97	500	595.65	0.84
29	99	538	496.45	1.08
30	99	457	456.23	1.00
31	100	423	565.83	0.75
32	105	652	619.38	1.05
33	115	582	587.57	0.99
34	115	442	512.96	0.86
35	119	581	562.86	1.03
36	120	420	502.01	0.84
37	120	449	494.23	0.91
38	121	803	663.47	1.21
39	122	566	559.88	1.01
40	124	581	717.61	0.81
41	126	486	670.78	0.72
42	127	675	588.61	1.15
43	128	473	554.57	0.85
44	130	722	788.34	0.92
45	135	535	535.00	1.00
46	136	866	922.94	0.94
47	136	856	698.78	1.22
48	137	521	479.81	1.09
49	138	695	785.05	0.89
50	139	423	587.93	0.72
51	139	1173	1036.87	1.13
52	139	707	916.88	0.77
53	141	641	589.68	1.09
54	144	583	598.85	0.97
55	149	934	953.43	0.98
56	151	621	549.51	1.13
57	151	685	638.10	1.07
58	155	598	550.83	1.09
59	157	618	654.31	0.94

UnitCode	# of admissions	Sum of observed LoS	Sum of predicted LoS	SLOSR
60	158	567	727.13	0.78
61	158	515	732.83	0.70
62	161	903	897.35	1.01
63	161	556	509.53	1.09
64	164	588	643.99	0.91
65	171	540	647.35	0.83
66	172	782	745.71	1.05
67	174	604	659.13	0.92
68	174	772	746.02	1.03
69	175	719	674.11	1.07
70	183	896	977.51	0.92
71	187	921	886.75	1.04
72	189	1035	902.23	1.15
73	195	723	757.19	0.95
74	197	1250	1181.51	1.06
75	199	985	909.52	1.08
76	200	1626	1197.54	1.36
77	211	717	1000.41	0.72
78	219	922	976.05	0.94
79	219	1043	895.24	1.17
80	220	1086	911.04	1.19
81	223	914	1024.14	0.89
82	223	939	1222.10	0.77
83	234	1174	1156.00	1.02
84	237	1030	1037.86	0.99
85	241	560	798.19	0.70
86	245	1072	920.19	1.16
87	245	957	1079.26	0.89
88	252	696	901.70	0.77
89	254	853	1047.48	0.81

UnitCode	# of admissions	Sum of observed LoS	Sum of predicted LoS	SLOSR
90	261	1016	1074.45	0.95
91	269	1355	1342.81	1.01
92	294	1254	1256.59	1.00
93	297	1572	1645.13	0.96
94	300	1429	1378.45	1.04
95	317	1022	1290.32	0.79
96	337	1353	1526.98	0.89
97	339	896	1131.82	0.79
98	340	1112	1109.25	1.00
99	353	1943	1436.04	1.35
100	362	1869	1684.90	1.11
101	373	1328	1571.65	0.84
102	374	1674	2164.13	0.77
103	422	2161	2149.24	1.01
104	425	1893	1723.31	1.10
105	445	1796	1859.89	0.97
106	502	1641	1662.66	0.99
107	539	2666	2120.18	1.26
108	697	2690	2623.42	1.03
109	757	3220	3094.52	1.04

Table C.1: Standardized Length of StayRatio (SLOSR) for each ICU

UnitCode	Average Age	Average Glasgow	Proportion of Mechanical Ventilation	Proportion of Clinical patients	Proportion of Elective Surgery patients	Proportion of Urgent Surgery patients	Mean ICU LoS
1	78.1	14.8	11.1%	72.2%	27.8%	0.0%	5.1
2	61.9	14.6	3.4%	93.1%	0.0%	6.9%	4.9
3	70.7	14.0	6.7%	93.3%	6.7%	0.0%	4.7
4	73.7	11.8	29.7%	97.3%	2.7%	0.0%	9.4
5	70.2	14.5	0.0%	71.8%	17.9%	10.3%	4.9
6	69.4	13.1	10.0%	52.5%	30.0%	17.5%	3.6
7	54.1	14.8	0.0%	80.5%	17.1%	2.4%	3.1
8	65.5	14.3	4.9%	7.3%	70.7%	22.0%	2.7
9	65.2	14.7	0.0%	92.2%	3.9%	3.9%	8.7
10	61.5	13.8	11.5%	48.1%	50.0%	1.9%	3.7
11	55.2	14.9	0.0%	86.2%	12.1%	1.7%	2.9
12	71.5	13.9	4.2%	80.3%	15.5%	4.2%	6.7
13	63.0	13.7	9.9%	70.4%	16.9%	12.7%	5.8
14	67.3	14.6	2.7%	65.3%	29.3%	5.3%	4.1
15	77.9	12.0	30.7%	98.7%	1.3%	0.0%	11.5
16	66.1	14.4	11.8%	55.3%	40.8%	3.9%	6.0
17	76.1	14.1	4.9%	92.6%	4.9%	2.5%	5.8
18	58.9	14.7	3.7%	67.1%	26.8%	6.1%	2.3
19	72.4	13.9	8.0%	96.6%	3.4%	0.0%	5.1
20	63.7	13.9	9.2%	97.7%	2.3%	0.0%	5.5
21	62.5	13.9	10.3%	81.6%	13.8%	4.6%	4.8
22	76.0	13.8	10.2%	95.5%	2.3%	2.3%	5.9
23	70.1	14.5	4.4%	80.0%	16.7%	3.3%	7.6
24	66.6	14.3	4.3%	84.9%	12.9%	2.2%	4.4
25	76.9	14.0	0.0%	96.8%	2.1%	1.1%	10.9

UnitCode	Average Age	Average Glasgow	Proportion of Mechanical Ventilation	Proportion of Clinical patients	Proportion of Elective Surgery patients	Proportion of Urgent Surgery patients	Mean ICU LoS
26	60.7	14.5	3.2%	72.6%	21.1%	6.3%	4.8
27	64.7	14.5	2.1%	93.7%	1.1%	5.3%	4.6
28	76.9	14.0	3.1%	94.8%	0.0%	5.2%	5.2
29	69.6	14.3	9.1%	69.7%	30.3%	0.0%	5.4
30	52.1	14.7	2.0%	88.9%	1.0%	10.1%	4.6
31	71.6	14.1	6.0%	89.0%	11.0%	0.0%	4.2
32	72.9	14.0	7.6%	94.3%	5.7%	0.0%	6.2
33	65.8	14.6	2.6%	100.0%	0.0%	0.0%	5.1
34	61.2	14.5	2.6%	95.7%	3.5%	0.9%	3.8
35	55.9	14.3	4.2%	85.7%	10.1%	4.2%	4.9
36	67.5	13.9	8.3%	72.5%	27.5%	0.0%	3.5
37	72.5	14.9	0.0%	75.8%	21.7%	2.5%	3.7
38	73.6	14.5	3.3%	78.5%	14.9%	6.6%	6.6
39	64.1	14.4	3.3%	83.6%	16.4%	0.0%	4.6
40	74.4	14.2	6.5%	94.4%	1.6%	4.0%	4.7
41	68.3	14.2	6.3%	67.5%	24.6%	7.9%	3.9
42	65.4	14.7	0.0%	83.5%	16.5%	0.0%	5.3
43	67.5	14.8	0.8%	85.9%	12.5%	1.6%	3.7
44	69.9	14.1	6.2%	92.3%	7.7%	0.0%	5.6
45	59.2	14.8	2.2%	54.1%	20.7%	25.2%	4.0
46	70.3	13.6	16.2%	94.9%	2.2%	2.9%	6.4
47	61.8	14.3	4.4%	99.3%	0.7%	0.0%	6.3
48	51.1	14.9	0.7%	90.5%	5.8%	3.6%	3.8
49	69.3	13.9	7.2%	89.1%	10.9%	0.0%	5.0
50	67.1	14.7	2.9%	70.5%	22.3%	7.2%	3.0
51	69.5	12.6	15.8%	98.6%	0.7%	0.7%	8.4
52	73.1	13.7	16.5%	93.5%	2.9%	3.6%	5.1
53	60.9	14.4	5.7%	46.8%	42.6%	10.6%	4.5
54	62.6	14.8	1.4%	88.2%	6.9%	4.9%	4.0
55	70.3	14.8	3.4%	90.6%	5.4%	4.0%	6.3

UnitCode	Average Age	Average Glasgow	Proportion of Mechanical Ventilation	Proportion of Clinical patients	Proportion of Elective Surgery patients	Proportion of Urgent Surgery patients	Mean ICU LoS
56	59.3	14.7	2.6%	73.5%	25.2%	1.3%	4.1
57	65.3	14.6	6.6%	64.2%	34.4%	1.3%	4.5
58	56.2	14.2	7.1%	20.6%	71.6%	7.7%	3.9
59	54.9	14.6	2.5%	86.6%	13.4%	0.0%	3.9
60	67.1	14.1	18.4%	63.3%	34.2%	2.5%	3.6
61	57.7	14.5	5.7%	72.8%	17.7%	9.5%	3.3
62	74.4	14.5	0.6%	78.3%	11.8%	9.9%	5.6
63	58.6	14.4	5.0%	8.7%	85.1%	6.2%	3.5
64	63.0	14.8	3.0%	90.2%	3.0%	6.7%	3.6
65	53.3	14.2	7.6%	28.1%	69.6%	2.3%	3.2
66	52.9	14.7	2.3%	93.6%	4.1%	2.3%	4.5
67	68.6	15.0	0.0%	79.9%	20.1%	0.0%	3.5
68	65.0	14.7	4.0%	89.7%	9.2%	1.1%	4.4
69	60.7	14.7	1.7%	84.0%	16.0%	0.0%	4.1
70	60.1	14.0	6.6%	97.3%	0.0%	2.7%	4.9
71	61.0	14.3	5.3%	84.5%	11.8%	3.7%	4.9
72	63.7	14.6	1.1%	97.4%	2.6%	0.0%	5.5
73	60.5	14.8	0.0%	67.7%	27.2%	5.1%	3.7
74	68.5	13.8	7.1%	99.5%	0.0%	0.5%	6.3
75	71.2	14.6	1.5%	75.4%	12.1%	12.6%	4.9
76	67.6	13.6	11.0%	78.5%	13.5%	8.0%	8.1
77	68.0	14.3	4.3%	62.6%	27.5%	10.0%	3.4
78	56.5	14.2	5.5%	71.7%	21.9%	6.4%	4.2
79	50.7	14.7	4.6%	93.6%	3.7%	2.7%	4.8
80	66.0	14.7	1.4%	66.4%	25.5%	8.2%	4.9
81	56.4	14.5	2.7%	77.6%	6.7%	15.7%	4.1
82	72.6	14.4	6.7%	80.7%	17.5%	1.8%	4.2
83	63.2	14.3	3.8%	97.4%	2.1%	0.4%	5.0
84	71.3	14.8	0.8%	87.8%	11.8%	0.4%	4.3
85	64.6	14.6	3.7%	16.2%	73.4%	10.4%	2.3

UnitCode	Average Age	Average Glasgow	Proportion of Mechanical Ventilation	Proportion of Clinical patients	Proportion of Elective Surgery patients	Proportion of Urgent Surgery patients	Mean ICU LoS
86	56.6	14.9	0.8%	90.2%	6.9%	2.9%	4.4
87	61.8	14.6	4.5%	96.7%	2.4%	0.8%	3.9
88	64.6	14.7	2.4%	57.1%	37.7%	5.2%	2.8
89	65.4	14.8	0.4%	95.7%	3.5%	0.8%	3.4
90	64.5	14.7	3.8%	24.5%	56.7%	18.8%	3.9
91	61.1	14.2	4.8%	98.1%	0.7%	1.1%	5.0
92	67.5	14.5	6.5%	78.6%	13.3%	8.2%	4.3
93	64.8	13.9	8.8%	81.8%	15.5%	2.7%	5.3
94	61.6	14.3	2.3%	94.0%	5.3%	0.7%	4.8
95	57.4	14.6	2.8%	81.4%	18.6%	0.0%	3.2
96	58.4	14.4	2.4%	90.5%	5.9%	3.6%	4.0
97	56.5	14.8	3.2%	63.1%	33.9%	2.9%	2.6
98	56.2	14.8	1.8%	92.1%	5.9%	2.1%	3.3
99	57.7	14.6	1.4%	89.8%	7.6%	2.5%	5.5
100	62.8	14.6	3.6%	87.3%	9.1%	3.6%	5.2
101	59.5	14.7	2.1%	96.2%	3.8%	0.0%	3.6
102	59.6	13.6	9.1%	82.6%	7.8%	9.6%	4.5
103	62.2	14.3	5.7%	96.7%	2.1%	1.2%	5.1
104	55.8	14.4	3.3%	91.5%	6.1%	2.4%	4.5
105	58.9	14.5	4.0%	81.3%	12.8%	5.8%	4.0
106	49.9	14.9	1.6%	89.8%	9.2%	1.0%	3.3
107	47.0	14.6	2.4%	94.4%	1.9%	3.7%	4.9
108	50.4	14.7	2.4%	91.8%	7.3%	0.9%	3.9
109	53.2	14.6	1.1%	69.6%	28.8%	1.6%	4.3

Table C.2: General description of each ICU