

# Série dos Seminários de Acompanhamento à Pesquisa

**DEI**  
DEPARTAMENTO  
DE ENGENHARIA  
INDUSTRIAL

Número 06 | 05 2021

## Prediction of Length of Stay in Intensive Care Units

Autor(es):  
Igor Tona Peres



## Prediction of Length of Stay in Intensive Care Units

Autor(es):

Igor Tona Peres

### CRÉDITOS:

SISTEMA MAXWELL / LAMBDA  
<https://www.maxwell.vrac.puc-rio.br/>

**Organizadores:** Fernanda Baião / Soraida Aguilar

**Layout da Capa:** Aline Magalhães dos Santos

Igor Tona Peres

PhD Student

Advisor: Fernando Luiz Cyrino Oliveira

Co-advisor: Silvio Hamacher

External advisor: Fernando Bozza

External advisor: Jorge Salluh

# Introduction

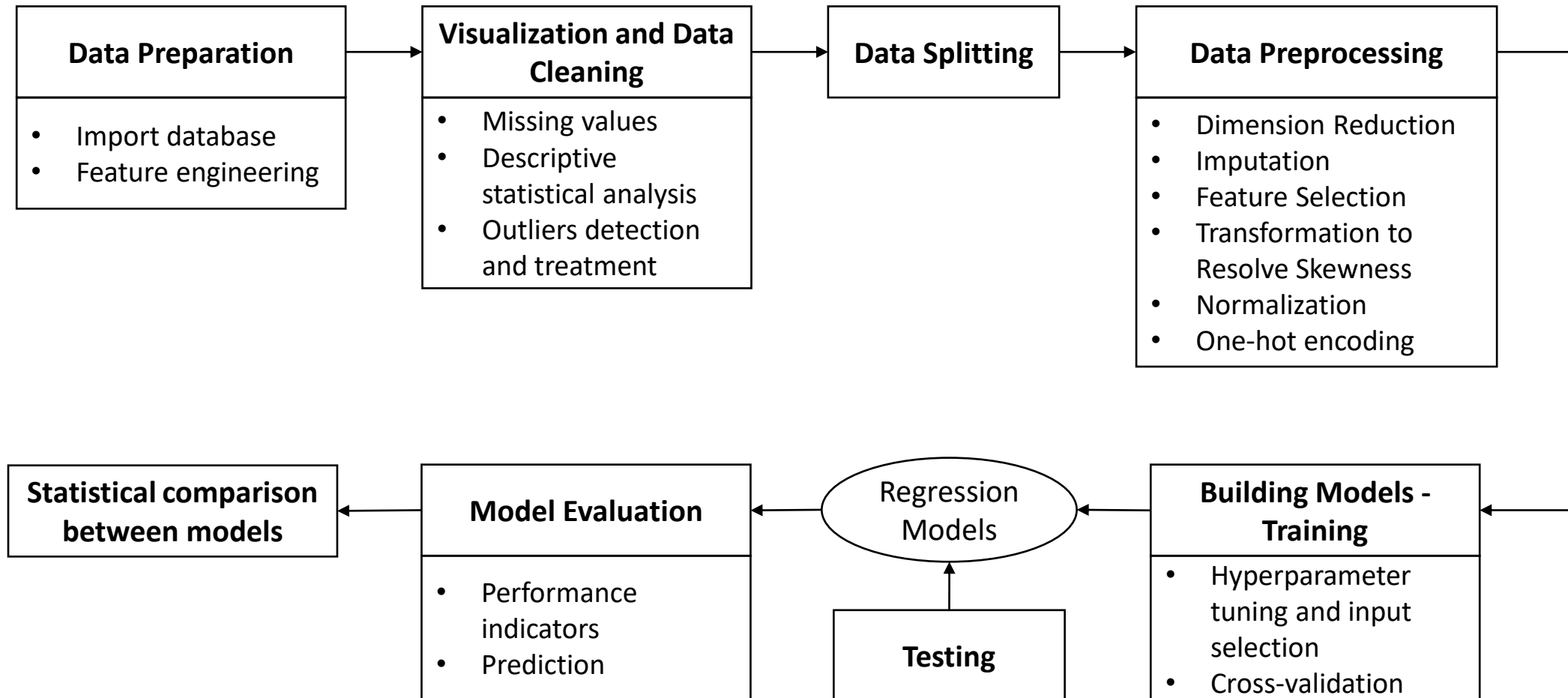
- Studies have shown that a small percentage of ICU patients presented a prolonged LoS.
- However, those few patients account for a large proportion of ICU days.
- Therefore, the early identification of prolonged stay patients can assist in improving unit efficiency.
- The main reasons for hospital administrators 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;
  - (iii) enabling case-mix adjustments for benchmarking analysis.

# Objectives

- To present a methodology that makes possible to predict the patient's length of stay in the Intensive Care Unit.
- Specific goals:
  - To propose and apply a data-driven methodology to predict the ICU LoS at day one
  - To dynamic adjust the prediction for the next seven days

# Methodology

- Figure 1 presents the framework of the proposed methodology, which was adapted from Dantas et al. (2020):



# Dataset description

- The data represents a set of 113 mixed-type ICUs from 40 different Brazilian hospitals.
- Total of 100,245 independent admissions from January 01 to December 31, 2019.
- The complete dataset is a join of six tables:
  - Sheet 2: Demographic data
  - Sheet 3: Comorbidities
  - Sheet 5: ICU complications (“first 24h”)
  - Sheet 6: Physiological and laboratory data (“first 1h”)
  - Sheet 8: Secondary diagnosis

# Dataset description

- Initial dataset = 103,195 independent admissions
- Final dataset = 100,340 independent admissions
- Filters:
  - Patients  $\geq 16$  and  $\leq 115$  years old (518 removed)
  - ICU LoS  $\geq 0$ h (108 removed)
  - ICU LoS  $\geq 6$ h (858 removed)
  - Previous hospital LoS lower than 60 days (876 removed)
  - Unit admission date  $\geq$  Hospital admission date (7 removed)
  - Presenting the admission main diagnosis code (1336 removed)



# Data Preparation

- Feature engineering

Adjusting existing features:

- Adjust the hospital length of stay prior to unit admission (using the dates)
- When ChronicHealthStatusName has value, all not informed comorbidities can be replaced by “No”
- Laboratorial and physiological features were changed
- Gender “undefined” was replaced by “Not informed” and then will be imputed.
- Admission source was reclassified
  - Surgical center; Ward/Room/Semi-intensive Unit; Home-care/Transfers/Others

# Data Preparation

- Feature engineering

Excluding organizational features

- ICU Type
- UnitCode
- HospitalCode
- Beds
- HealthInsuranceName

# Data Preparation

- Feature engineering

Propose new features from existing ones:

- Number of first day complications (“n\_complication”)
- Presence of any first day complication (“has\_complication”)

Treatment for “Secondary Diagnosis”

- Only diagnosis with more than 400 registries

# Visualization and Data Cleaning

- Missing values
  - Analyze the behavior of the missing data.
  - Features with more than 30% of missing will be excluded from the analysis (White et al., 2011).
  - Data with less than 30% of missing will be imputed.

# Results – Missing Data

- Features with more than 30% of missing:
  - Demographic data
    - ICDCode (100%);
  - Physiological and laboratory data (“first 24h”)
    - LowestPaCO21h (94%); LowestPaO21h (94%); LowestFiO21h (92%); HighestPaO21h (75%); HighestPaCO21h (75%); HighestFiO21h (61%);
  - Some relevant features will not be removed now:
    - PaO2FiO2 (83.4%); PH (74.5%); Lactate (65%); Bilirubin (57.5%); BMI (41%)

# Visualization and Data Cleaning

- Descriptive statistical analysis
  - We will apply a univariate analysis between the explicative variables and the ICU LoS.
    - Pearson Correlation for numerical variables
    - Cramer's V for categorical ones

# Results – Descriptive analysis

- Correlation with LoS for numeric variables

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

# Results – Descriptive analysis

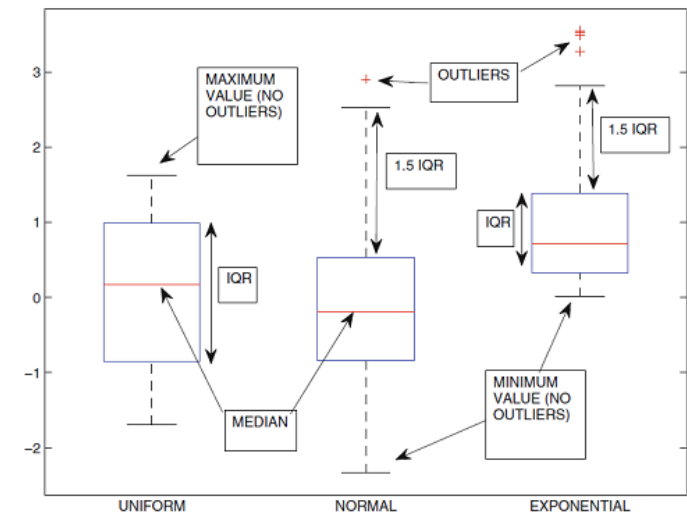
- Correlation with LoS for categorical variables

<b>Feature</b>	<b>Correlation</b>
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
IsChronicAtrialFibrillation	0.08
IsRenalReplacementTherapy	0.08
IsAcuteKidneyInjury	0.08



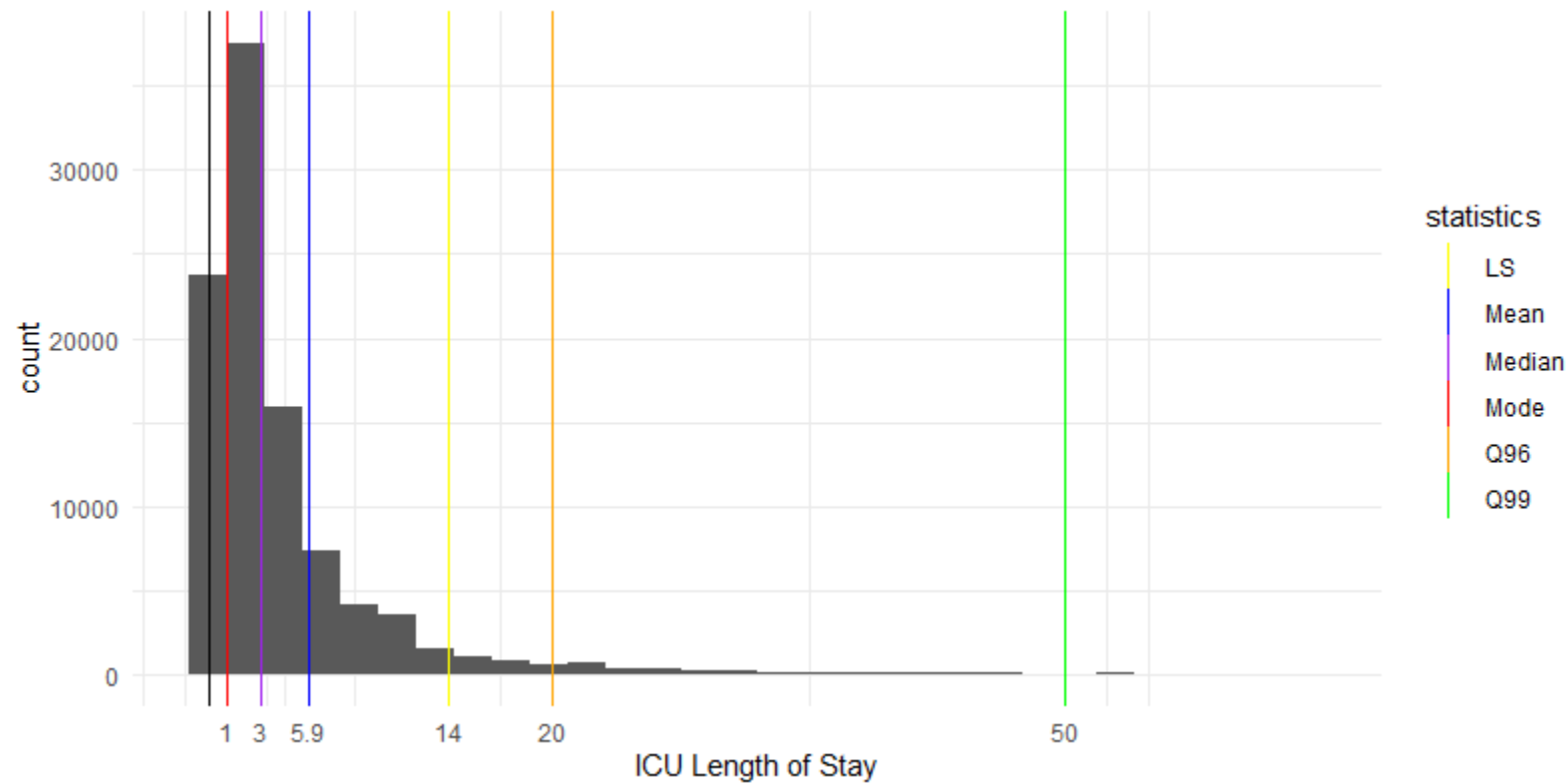
# Outlier detection and treatment

- Dispersion/Boxplot analysis
- 376 values were considered outliers
- Outliers will be replaced by “NA” and then imputed



# Outlier detection and treatment

- Histogram of ICU LoS



# Data Splitting

- We will apply this sampling methodology, splitting 80% of the dataset for training and 20% for testing
- To provide an unbiased sense of model effectiveness, the predictive model should be evaluated on samples that were not used to build or tune the model.
- Training with 5-fold cross-validation

# Data Preparation

- Feature engineering
  - Grouping “Admission Main Diagnosis” (851 possible classes)

Main Diagnosis Name	Total	%	Cumulative %	Mean ICU LoS	Median ICU LoS	Q90 ICU LoS	ICU Mortality
Pneumonia comunitária	6941	8.7%	8.7%	8.4	4.0	18.0	10%
Dor torácica	4188	5.2%	13.9%	2.6	2.0	5.0	0%
Infecção urinária sintomática, não especificada	3387	4.2%	18.1%	6.7	4.0	13.0	6%
Síncope	2983	3.7%	21.8%	3.7	2.0	6.0	1%
Angina instável	2592	3.2%	25.1%	3.3	2.0	6.0	0%
Insuficiência cardíaca aguda (descompensada)	2112	2.6%	27.7%	7.6	5.0	15.9	7%
AVC isquêmico	1651	2.1%	29.7%	7.7	3.0	16.0	5%
Epilepsia e transtornos convulsivos	1599	2.0%	31.7%	5.2	3.0	10.0	3%
Fibrilação atrial	1351	1.7%	33.4%	4.0	2.0	8.0	3%
Infarto miocárdico sem supra de ST	1211	1.5%	34.9%	5.2	3.0	10.0	4%
Tromboembolismo pulmonar	1155	1.4%	36.4%	5.0	3.0	8.0	3%
Intoxicações exógenas	1007	1.3%	37.6%	2.5	2.0	5.0	0%
Gastroplastias	996	1.2%	38.9%	1.3	1.0	2.0	0%
Acidente isquêmico transitório	988	1.2%	40.1%	3.0	2.0	5.0	1%
Gastroenterites / gastroenterocolites	954	1.2%	41.3%	3.7	2.0	7.0	2%
Outros diagnósticos, não classificados	924	1.2%	42.4%	4.0	3.0	8.0	3%
Hemorragia digestiva alta	894	1.1%	43.6%	5.2	3.0	10.0	5%
Outras complicações neurológicas	854	1.1%	44.6%	4.6	2.0	8.0	3%
DPOC descompensada	853	1.1%	45.7%	7.8	5.0	16.0	8%

# Data Preparation

- Feature engineering
  - Grouping “Admission - Main Diagnosis” (19 possible classes)

Main Diagnosis Name	Total	%	Cumulative %	Mean ICU LoS	Median ICU LoS	Q90 ICU LoS	ICU Mortality
Pneumonia comunitária	6941	8.7%	8.7%	8.4	4.0	18.0	10%
Dor torácica	4188	5.2%	13.9%	2.6	2.0	5.0	0%
Infecção urinária sintomática, não especificada	3387	4.2%	18.1%	6.7	4.0	13.0	6%
Síncope	2983	3.7%	21.8%	3.7	2.0	6.0	1%
Angina instável	2592	3.2%	25.1%	3.3	2.0	6.0	0%
Insuficiência cardíaca aguda (descompensada)	2112	2.6%	27.7%	7.6	5.0	15.9	7%
AVC isquêmico	1651	2.1%	29.7%	7.7	3.0	16.0	5%
Epilepsia e transtornos convulsivos	1599	2.0%	31.7%	5.2	3.0	10.0	3%
Fibrilação atrial	1351	1.7%	33.4%	4.0	2.0	8.0	3%
Infarto miocárdico sem supra de ST	1211	1.5%	34.9%	5.2	3.0	10.0	4%
Tromboembolismo pulmonar	1155	1.4%	36.4%	5.0	3.0	8.0	3%
Diagnosis Code with ICU LoS [0,1]	5419	6.8%	43.1%	2.0	1.0	3.0	0%
Diagnosis Code with ICU LoS ]1,2]	15680	19.6%	62.7%	3.5	2.0	6.0	2%
Diagnosis Code with ICU LoS ]2,3]	14063	17.5%	80.2%	5.0	3.0	10.0	4%
Diagnosis Code with ICU LoS ]3,4]	8777	10.9%	91.2%	7.6	4.0	15.0	9%
Diagnosis Code with ICU LoS ]4,5]	2589	3.2%	94.4%	8.6	5.0	18.0	9%
Diagnosis Code with ICU LoS ]5,7]	1389	1.7%	96.1%	13.3	7.0	31.0	18%
Diagnosis Code with ICU LoS >7	114	0.1%	96.3%	14.0	10.0	33.0	46%
Others	2996	3.7%	100.0%	6.0	3.0	14.0	7%

# Data Preparation

- Feature engineering
  - Grouping “Admission - Main Diagnosis” (9 possible classes)

Main Diagnosis Name	Total	%	Cumulative %	Mean ICU LoS	Median ICU LoS	Q90 ICU LoS	ICU Mortality
[0,1]	5419	6.8%	6.8%	2.0	1.0	3.0	0.0%
]1,2]	26794	33.4%	40.2%	3.4	2.0	6.0	1.0%
]2,3]	19679	24.5%	64.7%	5.3	3.0	10.0	4.0%
]3,4]	19105	23.8%	88.5%	7.7	4.0	16.0	9.0%
]4,5]	4701	5.9%	94.4%	8.1	5.0	17.0	8.0%
]5,6]	477	0.6%	95.0%	11.3	6.0	26.0	14.0%
]6,7]	912	1.1%	96.1%	14.3	7.0	34.9	20.0%
>7	114	0.1%	96.3%	14.0	10.0	33.0	46.0%
Others	2996	3.7%	100.0%	6.0	3.0	14.0	7.0%

# Data Preparation

- Feature engineering
  - Grouping “Admission - Main Diagnosis” (8 possible classes)

Main Diagnosis Name	Total	%	Cumulati ve %	Mean ICU LoS	Median ICU LoS	Q90 ICU LoS	ICU Mortality
[0,1]	1064	1.3%	1.3%	1.7	1.0	2.8	0.0%
]1,2]	16623	20.9%	22.2%	2.5	1.7	4.7	1.0%
]2,3]	27129	34.1%	56.3%	3.8	2.6	7.7	2.0%
]3,4]	17201	21.6%	77.9%	5.4	3.6	12.7	6.0%
]4,5]	14860	18.7%	96.6%	6.7	4.6	17.1	10.0%
]5,6]	1170	1.5%	98.0%	7.7	5.6	21.0	11.0%
]6,7]	334	0.4%	98.4%	8.5	6.5	21.0	14.0%
>7	1242	1.6%	100.0%	10.4	8.8	21.0	23.0%

# Data Preprocessing

- Transformations of the training dataset to improve model performance.
  - Dimension reduction
  - Imputation
  - Feature Selection
  - Transformation to Resolve Skewness
  - Normalization
  - One-hot encoding



# Data Preprocessing

- Dimension reduction
  - Zero and Near Zero Variance Predictors
    - Vast majority of cases presenting a unique value and few cases presenting other values
  - Identifying Correlated Predictors
    - redundant predictors can add more complexity to the model than information
    - for numeric variables, we used Pearson Correlation with a recommended threshold of 0.75
    - for categorical, we employed Cramér's V with a recommended threshold of 0.5

# Results – Dimension reduction

- Removing Zero and Near Zero Variance
  - Demographic data:
    - IsReadmission24h;IsReadmission48h
  - Comorbidities:
    - IsOtherSolidOrganTransplant
    - IsAtrialFlutter
    - IsCombinedPancreaskidneyTransplant
    - IsHyperthyroidism
    - IsAllogeneicBMT
    - IsAutologousBMT
    - IsPepticDisease
  - ICU complications:
    - IsNeutropenia
    - IsVentricularSustainedCardiopulmonary
    - IsCombinedLiverkidneyTransplant
  - Laboratorial data:
    - PaO2FiO2

# Results – Dimension reduction

- Removing Zero and Near Zero Variance

- Demographic data:

- IsReadmission24h;IsReadmission48h

- Comorbidities:

- IsOtherSolidOrganTransplant
    - IsAtrialFlutter
    - IsCombinedPancreaskidneyTransplant
    - IsHyperthyroidism
    - IsAllogeneicBMT
    - IsAutologousBMT →
    - IsPepticDisease

Feature	Class	Mean	SD	N	Beta	CI	Pvalue
IsAutologousBMT	0	5.35	10.83	80168	ref.	-	-
IsAutologousBMT	1	6.38	8.42	29	1.03	[-2.91, 4.97]	0.6094

- ICU complications:

- IsNeutropenia
    - IsVentricularSustainedCardiopulmonary
    - IsCombinedLiverkidneyTransplant

- Laboratorial data:

- PaO2FiO2

# Results – Dimension reduction

- Correlation for numeric features (collinearity analysis)

<b>Feature 1</b>	<b>Feature 2</b>	<b>Correlation</b>
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

- Removed features (threshold=0.75):
  - BUN; MFIpoints; Saps3DeathProbabilityStandardEquation; LowestDiastolicBloodPressure1h; LowestSystolicBloodPressure1h.

# Results – Dimension reduction

- Cramer-V for categorical features (collinearity analysis)

<b>Feature 1</b>	<b>Feature 2</b>	<b>Correlation</b>
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

- Removed features (threshold=0.47):  
IsCardiopulmonaryArrest; IsImmunossupression; AdmissionReasonName; has\_complication

# Data Preprocessing

- Imputation
  - Inadequate handling of missing data can lead to biased or inefficient estimates of parameters
  - Missing at random (MAR) — the probability of data being missing does not depend on the unobserved data, conditional on the observed data
    - Multiple imputation by chained equations (MICE)
  - Missing not at random (MNAR) — the probability of data being missing does depend on the unobserved data, conditional on the observed data
    - Laboratory variables: replacing missing values with the mean or median of the observed values

# Data Preprocessing

- Feature Selection
  - Reduce the dimension of the problem by removing variables that do not present significant contribution to the model.
    - Recursive Feature Elimination with random forest (RF-RFE),
    - LASSO
    - Selection by Filter (SBF)

# Data Preprocessing

- Transformation to Resolve Skewness
  - Right skewed distribution can affect the performance of some classifiers
  - Replacing that data with the log, square root, or inverse may help to remove the skew
  - Box-cox transformations use maximum likelihood estimation to determine lambda in training data

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

- Log transformation (lambda=0); square transformation (lambda = 2), square root (lambda = 0.5), inverse (lambda = -1), and others in-between.



# Data Preprocessing

- Normalization
  - Set the numerical variables of the database on a common scale.
  - Improve the numerical stability of some calculations.
  - To scale the data between 0 and 1
  - Min-Max Normalization or Normalization by Range:

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

# Data Preprocessing

- One-hot encoding
  - One-hot encoding is a method used to handle datasets with mixed data types (numerical, categorical and binary)
  - Some machine learning prediction models does 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.

# Over-Fitting and Model Tuning

- The hyper-parameters of the model will be selected in order to minimize the prediction error
- Measure of error:
  - Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

# Regression Models

- We will test the following machine learning techniques:
  - Support Vector Regression (SVR)
  - CART
  - Bagging
  - Gradient Boosting Trees
  - Random Forests
  - k-nearest Neighborhood
  - Linear Regression
  - Glm with Negative Binomial distribution

# Scenarios

- Main Diagnosis
  - 851 codes
  - 19 grouped codes
  - 9 grouped codes
  - 8 grouped codes
- ICU length of stay:
  - Raw
  - Truncated
  - Log-transformed
- Feature selection analysis
  - All features
  - Severity scores (MFI, SOFA, SAPS3) x Raw features (Comorbidities, Complications, and Laboratorial data)
  - Missing data analysis
- Comparison between regression models

# Results

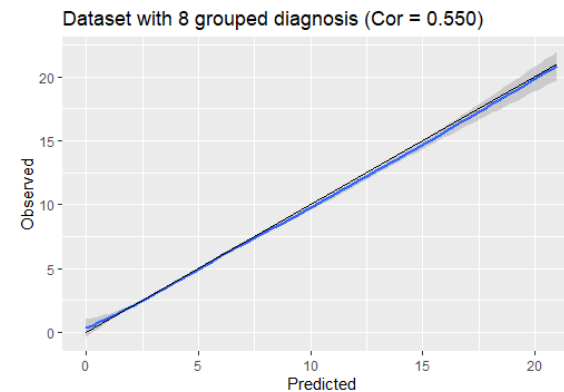
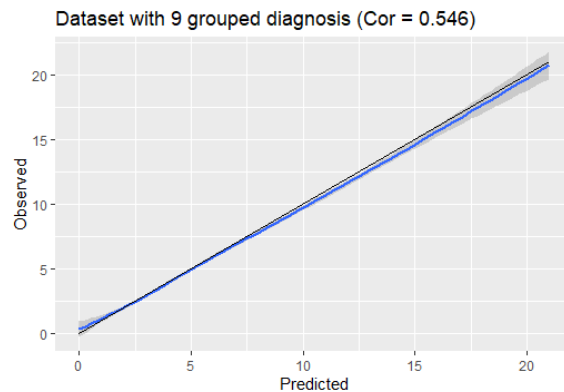
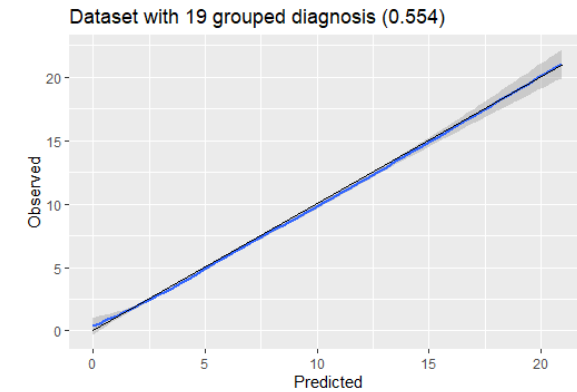
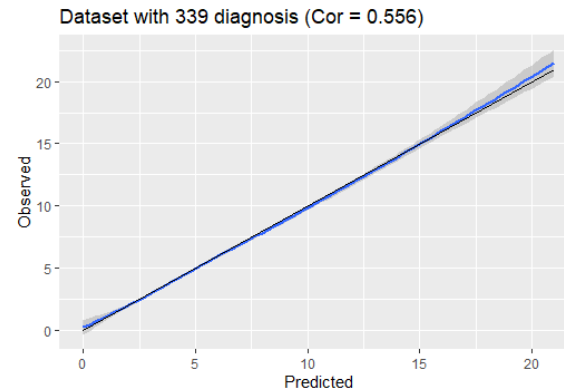
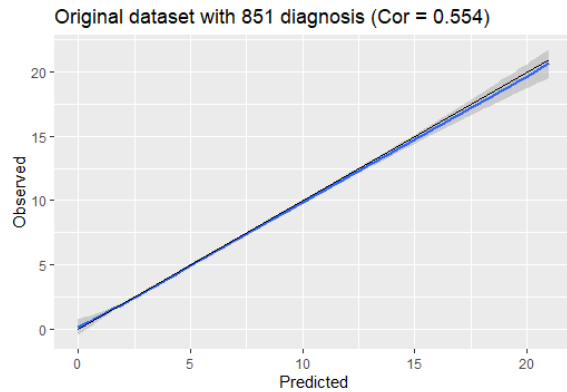
- Scenario 1: Comparison between “Main Diagnosis” groups
  - Truncated LoS; All features; No feature selection; model = “boosting”

Type of grouping for "Main Diagnosis"	Testing set				Training set					
	RMSE	MAE	R <sup>2</sup>	Cor	RMSE (SD)	MAE (SD)	R <sup>2</sup>	(SD)	(SD)	(SD)
Original dataset with 851 diagnosis codes	3.97	2.64	31%	0.554	3.96	0.06	2.60	0.03	33%	0.01
Dataset with 339 diagnosis codes (Diagnosis <20 = Others)	3.97	2.64	31%	0.556	3.96	0.03	2.60	0.01	32%	0.01
Dataset with 19 grouped diagnosis (Diagnosis <20 = Others)	3.97	2.64	31%	0.554	3.96	0.03	2.60	0.01	32%	0.00
Dataset with 9 grouped diagnosis (Diagnosis <20 = Others)	4.00	2.66	30%	0.546	3.97	0.04	2.60	0.01	32%	0.01
<b>Dataset with 8 grouped diagnosis</b>	<b>3.99</b>	<b>2.65</b>	<b>30%</b>	<b>0.550</b>	<b>3.96</b>	<b>0.03</b>	<b>2.59</b>	<b>0.01</b>	<b>32%</b>	<b>0.01</b>




# Results

- Scenario 1: Comparison between “Main Diagnosis” groups
  - Truncated LoS; All features; No feature selection; model = “boosting”



# Results

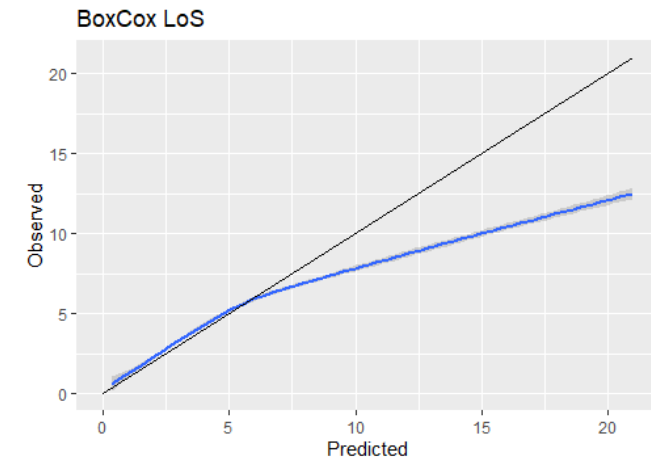
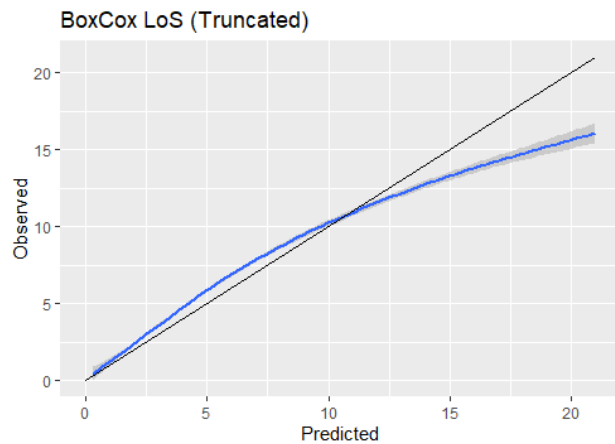
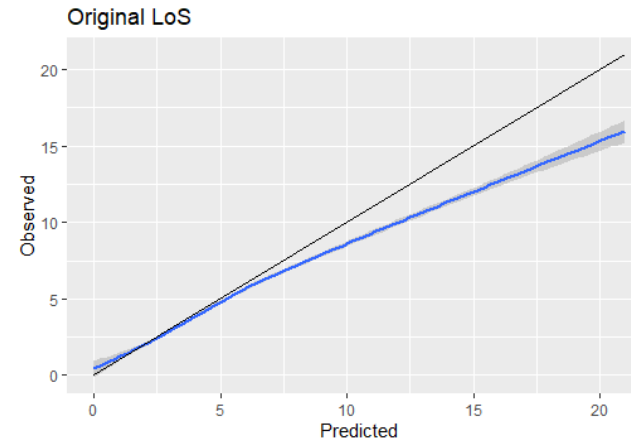
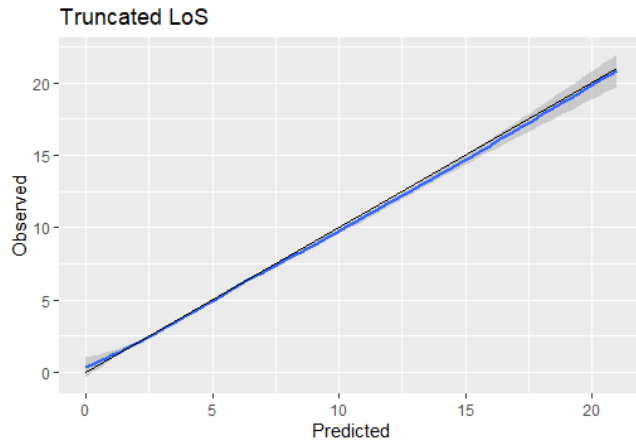
- Scenario 2: Comparison between types of LoS
  - All features; No feature selection; model = “boosting”; 8 diagnosis groups
  - All models were compared truncating the ICU LoS bigger than 21 days (in the test set)

Type of LoS	Testing set			
	RMSE	MAE	R <sup>2</sup>	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



# Results

- Scenario 2: Comparison between types of LoS
  - All features; No feature selection; model = “boosting”; 8 diagnosis groups



# Results

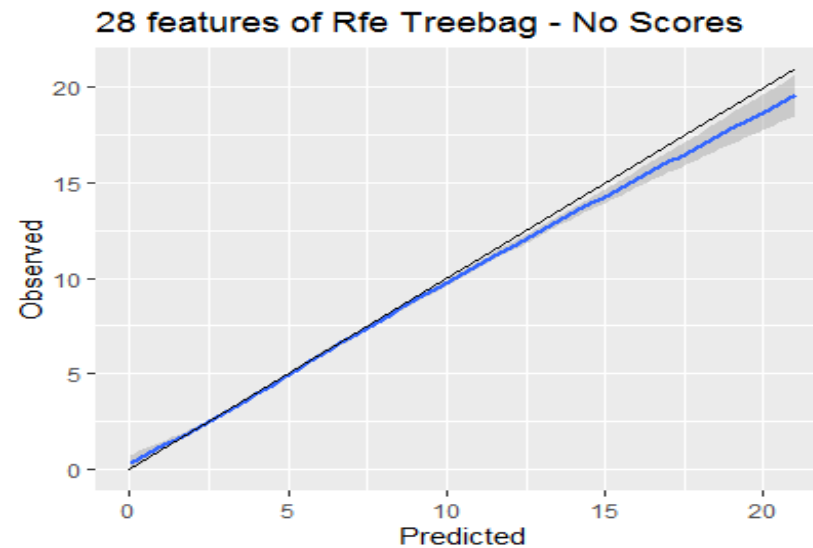
- Scenario 3: Severity scores, feature selection and missing data analyses
  - Truncated LoS; 8 diagnosis groups; model = “boosting”
  - Severity scores = SAPS3, SOFA, MFI, Charlson Comorbidity Index

Severity scores, feature selection and missing data analyses	Testing set				Training set					
	RMSE	MAE	R <sup>2</sup>	Cor	RMSE	(SD)	MAE	(SD)	R <sup>2</sup>	(SD)
Dataset with all features (90 features)	3.99	2.65	30%	0.550	3.96	0.03	2.59	0.01	32%	0.01
<b>Severity scores analysis</b>										
Dataset considering only severity scores features (4 features)	4.29	2.85	19%	0.440	4.32	0.02	2.86	0.01	20%	0.00
Dataset not considering severity scores features (86 features)	4.01	2.67	29%	0.544	3.97	0.04	2.61	0.03	32%	0.01
<b>Feature Selection analysis</b>										
Dataset after feature selection (14 features of Rfe Treebag)	4.08	2.69	27%	0.519	4.06	0.02	2.66	0.01	29%	0.00
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%	0.01
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%	0.00
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%	0.00
Dataset after feature selection without severity scores and glasgow (25 features of Rfe Treebag)	4.06	2.70	28%	0.525	4.02	0.02	2.64	0.01	30%	0.01
<b>Missing data analysis</b>										
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%	0.00
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%	0.01




# Results

- Scenario 3: Severity scores and feature selection analysis
  - Truncated LoS; 8 diagnosis groups; model = “boosting”



# Results

- Scenario 4: Comparison of Regression Models
  - Truncated LoS; 8 Diagnosis Groups; 28 selected features from RFE



Models	Testing set				Training set					
	RMSE	MAE	R <sup>2</sup>	Cor	RMSE (SD)	MAE (SD)	R <sup>2</sup>	(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

# Results – Variable Importance

- Random Forests

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%
IsAcuteAtrialFibrillation	0.87%
IsAlcoholism	0.85%
IsAcuteKidneyInjury	0.70%
IsAids	0.20%

# Conclusions

Benefits of obtaining a good estimate for the length of stay:

- Assistance level:
  - optimize the implementation of healthcare protocols (such as sedation and mobilization),
  - discussion of the cases in each multi-professional round,
  - better preparation of healthcare transit.
- Operational level:
  - planning the ICU discharge,
  - prioritizing patients to be evaluated daily,
  - better communication between family members, teams, and managers.
- Strategic level:
  - better sizing the number of beds,
  - improving the benchmarking analysis between ICUs.