Predicting ICU Length of Stay

for Patients with

Acute Traumatic Spinal Cord Injury

Using MIMIC Data

---

1. Introduction

Case Study Overview

The prediction of **ICU Length of Stay (LOS)** is a critical challenge in healthcare management, particularly for patients with **acute traumatic spinal cord injury (SCI)**. This injury often results in extended ICU stays due to complications such as respiratory failure, infections, and neurological deficits. In particular, **cervical spinal cord injuries** (ICD-9 codes: 806.0–806.09) often lead to more severe and complex medical conditions that prolong ICU stays.

This study leverages the **MIMIC-III** dataset, which contain detailed ICU admission records from patients at Beth Israel Deaconess Medical Center. These datasets include patient demographics, diagnostic codes, treatment interventions, and ICU discharge information. The study aims to predict the ICU length of stay more accurately than the traditional **APACHE-IV model**, which has shown limited predictive power (R^2 = 0.05–0.28) for trauma patients like those with SCI.

Research Question

The main research question of this study is:

- **Can we develop a model that predicts ICU length of stay (LOS) for SCI patients more accurately than the APACHE-IV model?**

The goal is to improve predictions of ICU LOS, which will help in better planning and resource allocation for SCI patients in the ICU.

---

2. Data Plan and Preparation

We analyzed each CSV file deeply and reported our findings in a Google doc.

- **Dataset Description**

This study utilizes the **MIMIC-III (Medical Information Mart for Intensive Care)** database, which contains de-identified health data for patients who were admitted to the intensive care units (ICU) at Beth Israel Deaconess Medical Center from 2001 to 2012. The dataset provides comprehensive information on patient demographics, medical history, diagnoses, treatments, clinical interventions, and patient outcomes. The data from MIMIC-III is organized across multiple tables that capture different aspects of patient care, including admissions, ICU stays, diagnosis codes, and recorded clinical events.

- **Data Dictionary :**

I. **`patients`**: Contains demographic information about patients, including their **`subject_id`** (patient identifier), **`age`**, **`gender`**, **`ethnicity`**, and other details such as **`marital_status`** and **`insurance`**.

II. **`admissions`**: Contains information about patient admissions to the hospital, including **`hadm_id`** (hospital admission identifier), **`admission_type`** (emergency, elective), **`admission_location`** (e.g., emergency room, transfer from another hospital), and **`hospital_expire_flag`** (whether the patient died in the hospital).

III. **`icustays`**: Contains data on patient stays in the ICU, including **`icustay_id`** (ICU stay identifier), **`los`** (ICU length of stay), **`intime`** (ICU admission time), **`outtime`** (ICU discharge time), and the type of ICU unit (e.g., MICU, SICU).

IV. **`diagnoses_icd`**: Contains the ICD-9 codes for diagnoses, which are key to identifying and categorizing conditions such as **spinal cord injuries** (SCI). Specific codes such as **806.0 - 806.09** (Cervical SCI) are used in this study.

V. **`chartevents`**: Contains clinical observations, including recorded vital signs such as **heart rate**, **blood pressure**, **temperature**, and **respiratory rate**.

VI. **`procedureevents_mv`**: Contains data related to mechanical ventilation use, including the **`ventilator_duration`**, an important factor in predicting ICU LOS.

VII. **`inputevents`**: Records medications, fluids, and other therapeutic interventions administered to patients during their ICU stay.

VIII. **`labevents`**: Contains lab results such as **blood gas analysis**, **glucose levels**, and **inflammatory markers**, which are crucial in understanding patient condition and recovery trajectory.

3. Data Preprocessing

- **Data Integration**

  The data was pulled from various tables within the MIMIC-III database:

  `ICUSTAYS.csv`, `ADMISSIONS.csv`, `DIAGNOSES_ICD.csv`, `CHARTEVENTS.csv`, `PROCEDUREEVENTS_MV.csv`, and others were merged using common identifiers like `subject_id`, `hadm_id`, and `icustay_id`.

- **Data Cleaning**

  I. **Missing Data**: Some critical columns, such as ICU discharge timestamps and diagnostic codes, had missing values. Missing timestamps were imputed where possible, and rows with missing `outtime` values were excluded from analysis.
  II. **Duplicates**: Duplicate records were identified based on `subject_id` and `hadm_id` and removed to ensure unique entries.
  III. **Standardization**: All dates were converted to a consistent format, and numeric features (e.g., `ventilator_duration`, `age`) were scaled to improve model performance.

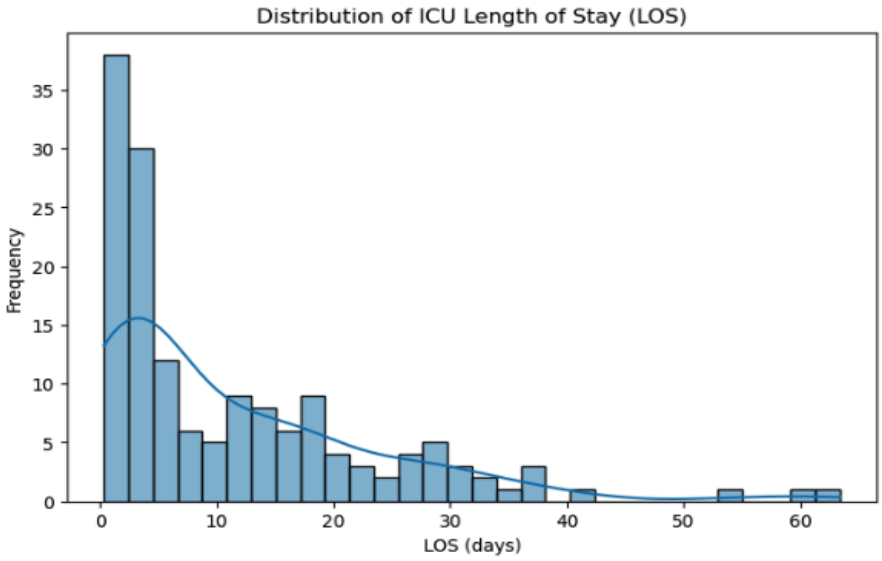
Feature Engineering

- The `ventilator_duration` was calculated from the `procedureevents_mv.csv` to track the duration of mechanical ventilation. This feature was crucial as prolonged ventilation is expected to correlate with longer ICU stays.
- **Categorical Features**: The `gender`, `ethnicity`, and `insurance` were converted into categorical variables using one-hot encoding to be used in machine learning models.

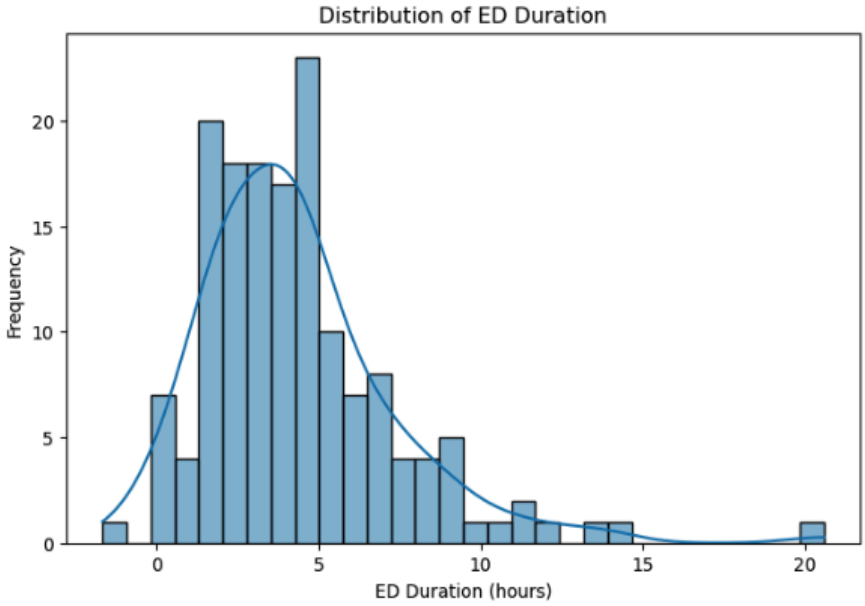
4. Exploratory Data Analysis (EDA)

Visualizations:

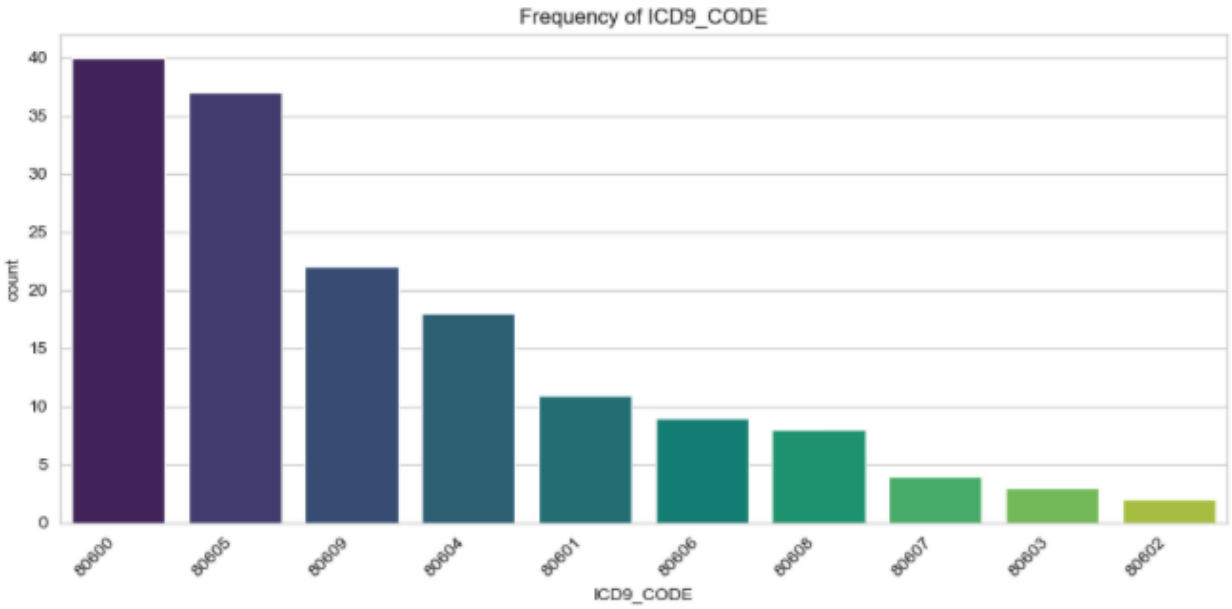
Distribution of ICU Length of Stay (LOS)



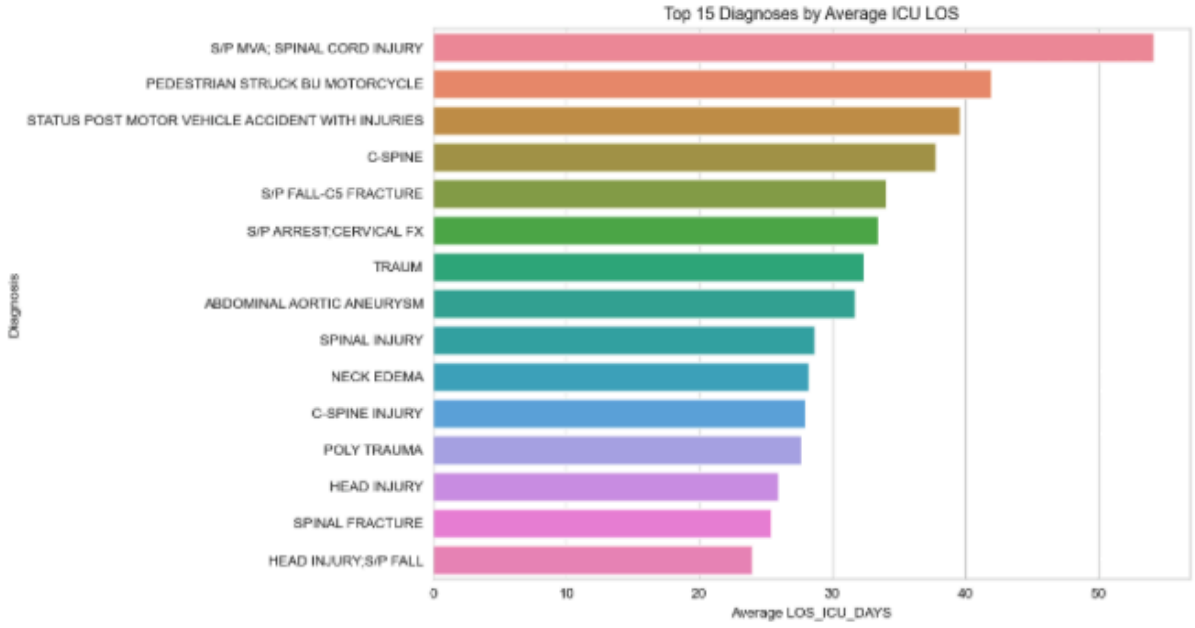
Distribution of ED Duration



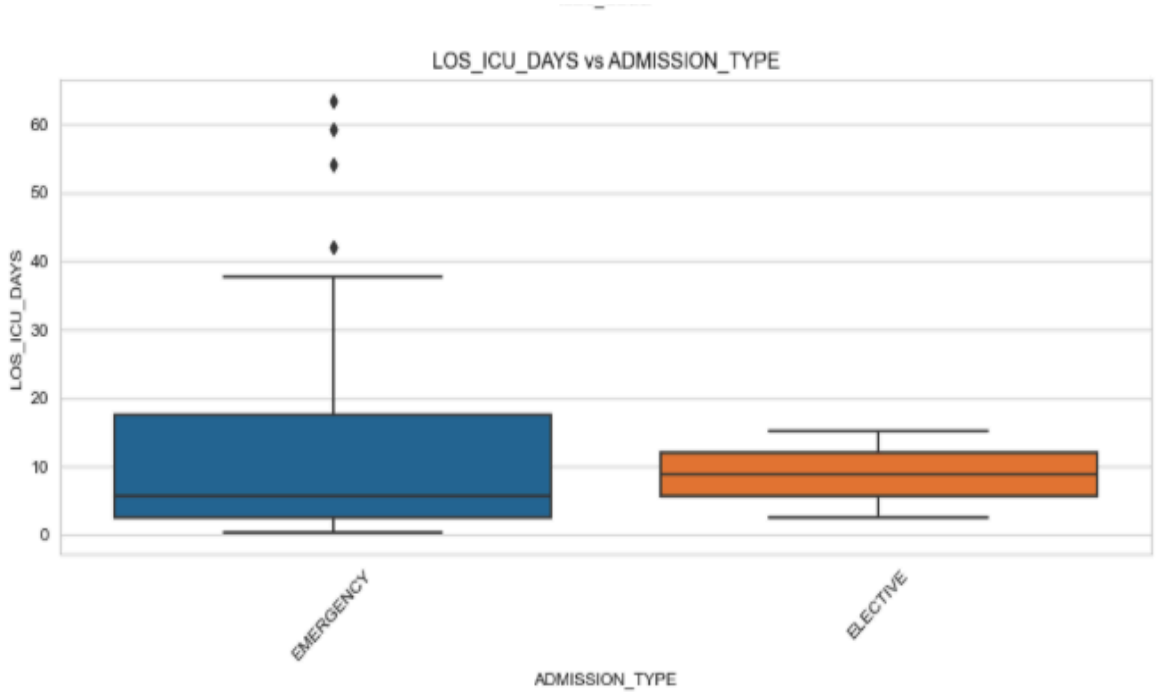
Frequency of ICD9 Codes

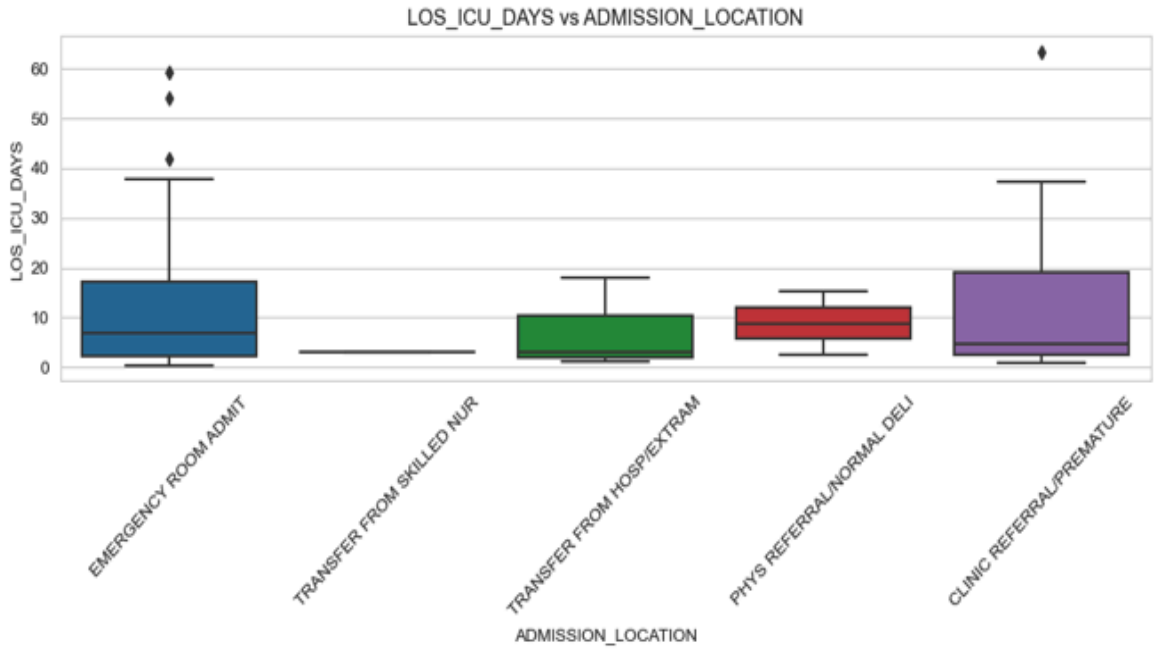


Top 15 Diagnoses by Average ICU LOS

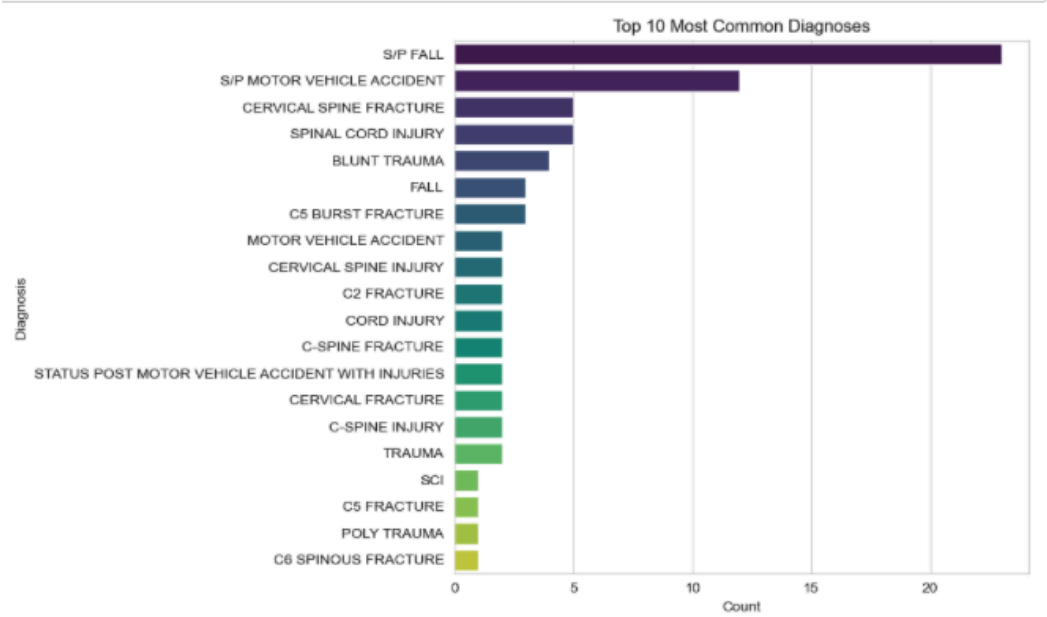


LOS vs. Admission Type & LOS vs. Admission Location

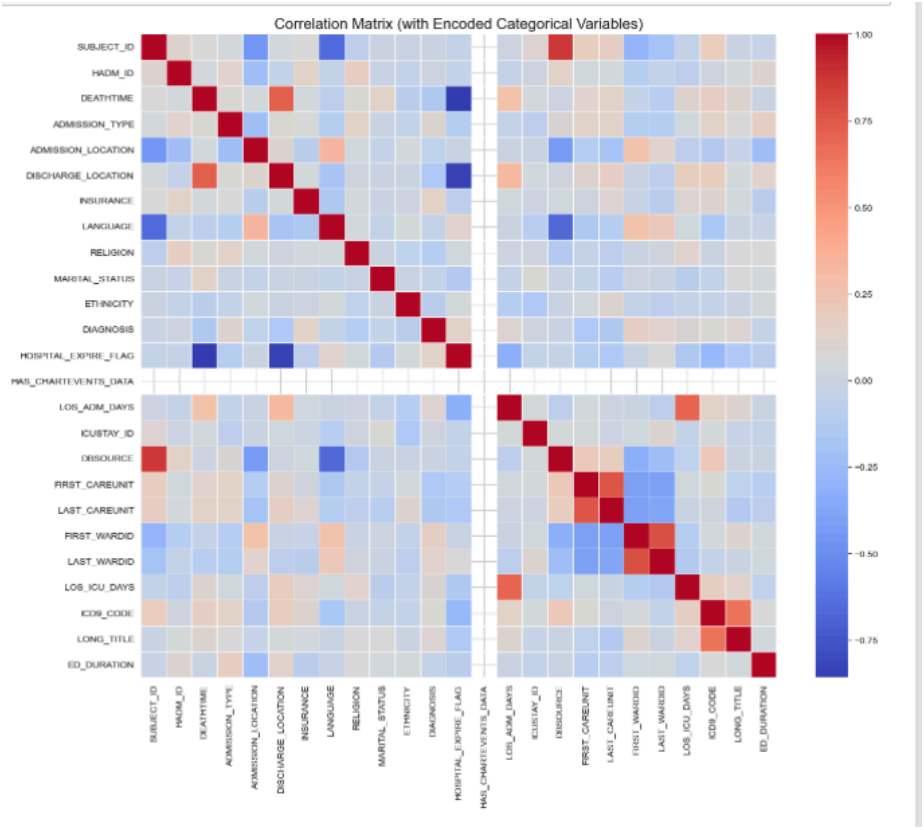




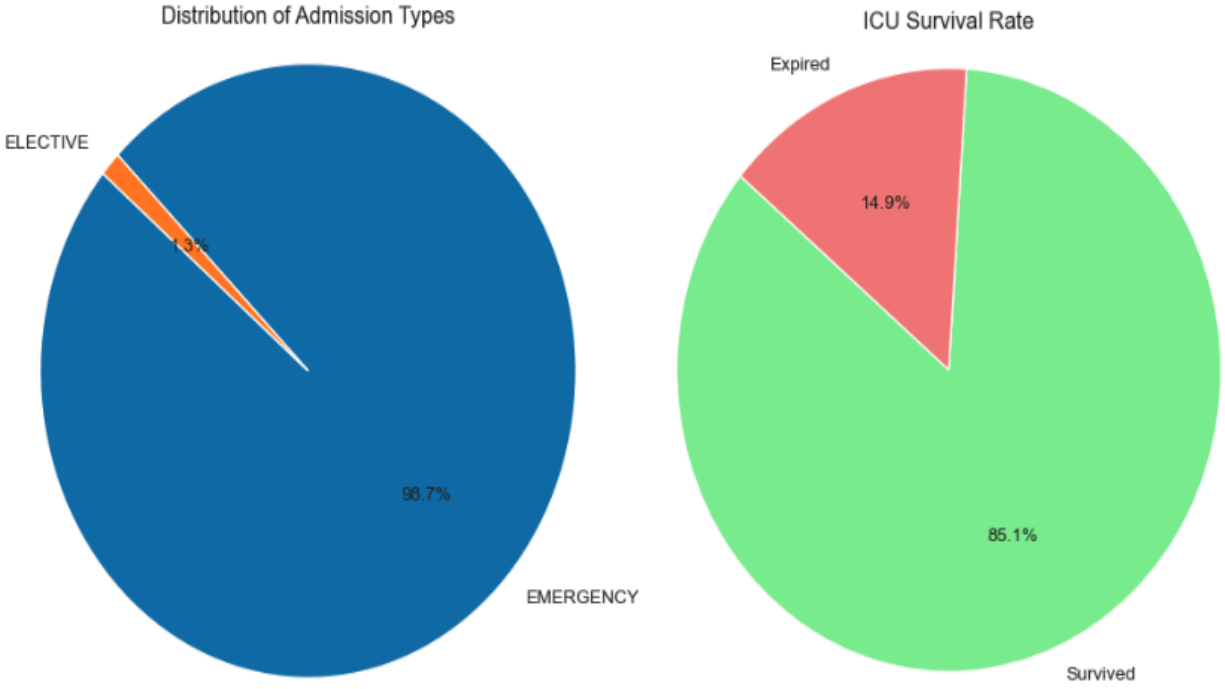
Top 10 Most Common Diagnoses



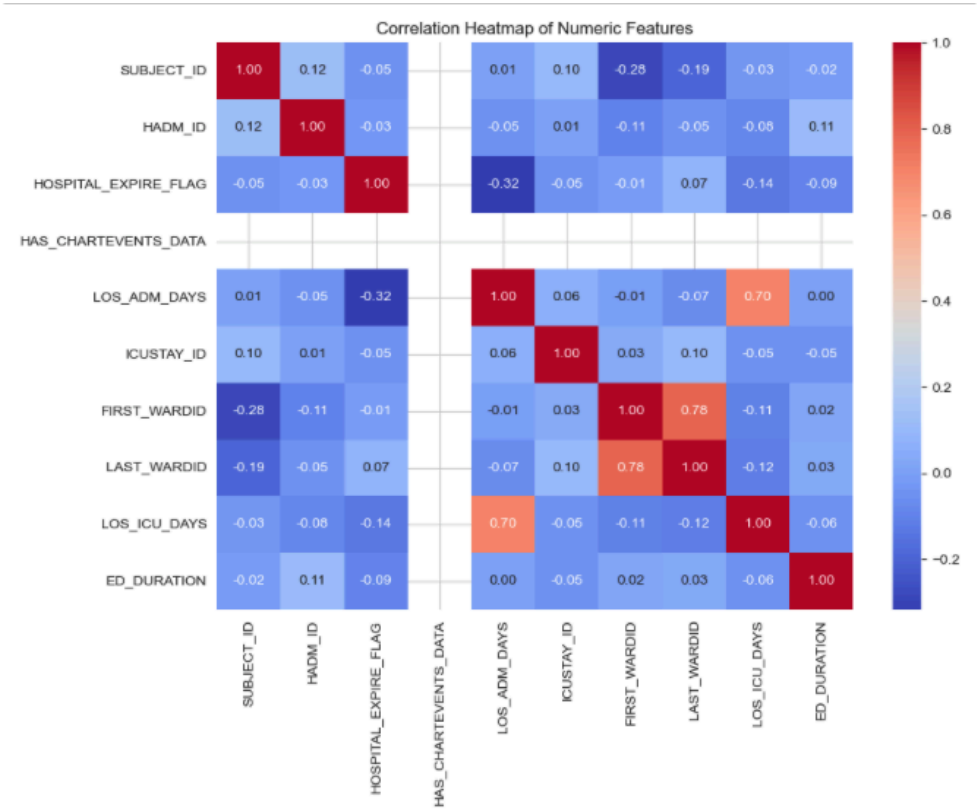
Correlation Matrix (Categorical Variables)



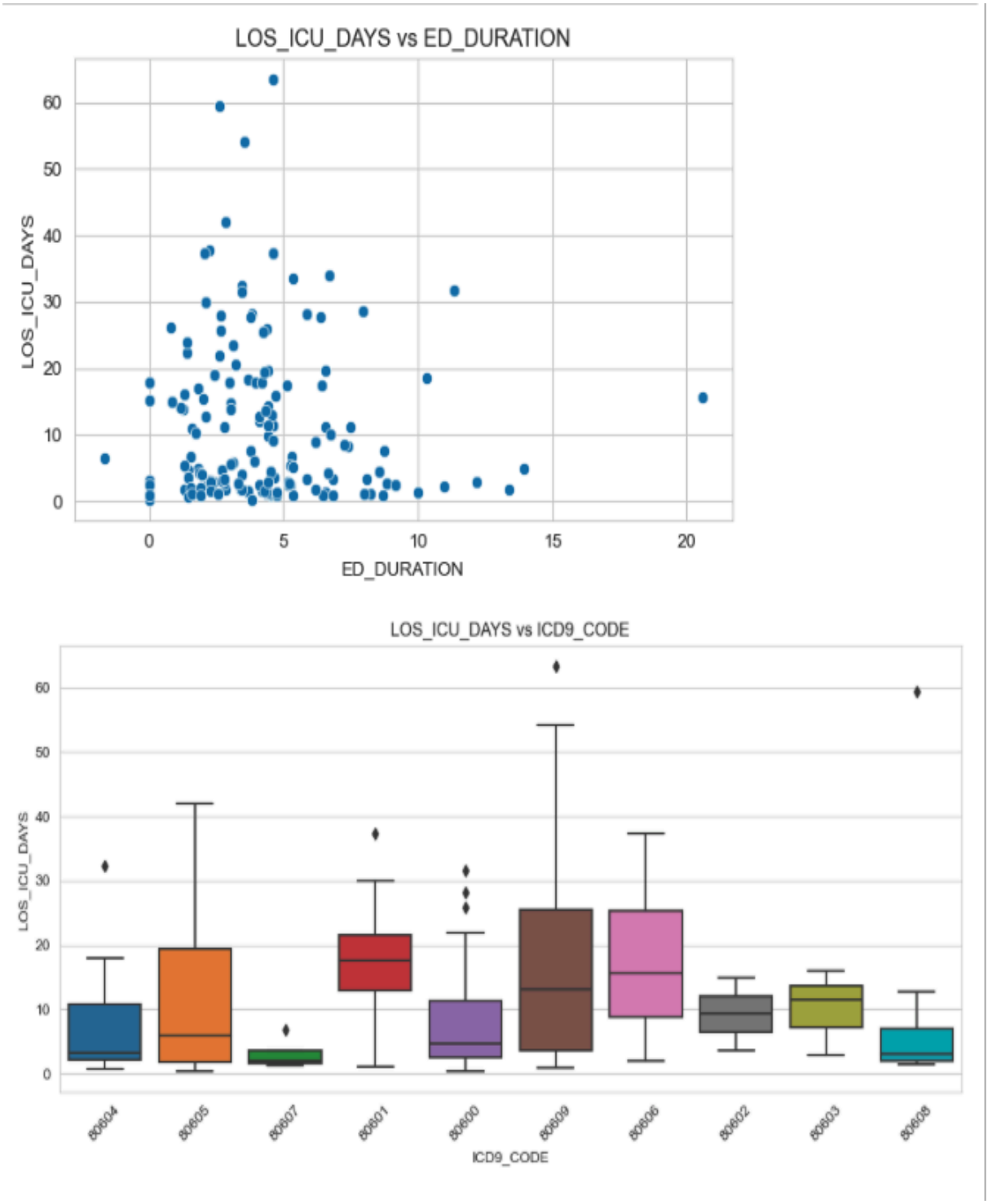
Distribution of Admission Types & ICU Survival Rates



Correlation Heatmap of Numeric Features



LOS_ICU_DAYS vs ED_DURATION & LOS_ICU_DAYS vs IC9_CODE



Total Patients: Only 171 patients have any form of 806.0x spinal cord injury.

Most Common Diagnosis: 806.00 (C1–C4 fx-cl/cord injury NOS) is the most common, with 49 cases.Age Group Most Affected: 50s and 70s age groups have the highest overall counts for these injuries.

Injuries in Young Adults: There are cases in the 20s and 30s, showing that traumatic spinal injuries also occur in younger adults (especially 806.05 and 806.06).Rare in Children and Elderly: 0s and 10s age groups have near-zero cases, and the 90+ group has only a few.

1. Primary vs. Secondary Role of Diagnosis

Most spinal injury cases are listed as primary diagnosis, suggesting they are the main reason for admission.Example: 806.00 has 35 primary vs. 14 secondary.806.09 is notable: 78% are primary, indicating high clinical significance.

2. Rare Cord Syndromes Often Primary

Even rare codes like:806.03 (C1–C4 central cord) — 100% primary,806.07 (C5–C7 anterior cord) — 100% primary

This implies that central or anterior cord injuries, though uncommon, are often the chief concern at admission.

3. Unspecified Injuries Most Frequent

Codes with “NOS” (Not Otherwise Specified):806.00 and 806.05 are the top 2 in total count. This suggests lack of detailed spinal cord classification in some clinical notes or imaging.

4. Injury Location Trend

More injuries were reported in C1–C4 than C5–C7. This could indicate that these higher cervical injuries are either more common or more likely to require ICU-level admission (which MIMIC-III captures).

5. High Primary Diagnosis Rate Indicates Severity

Across all spinal injury codes, ~80% are primary diagnoses.This confirms spinal trauma is not only serious but often the main clinical concern in ICU admissions.

100% primary diagnosis rate for:

806.03 (C1–C4 central cord syndrome) — most common in 50s

806.07 (C5–C7 anterior cord syndrome) — peaks in 20s

High severity (≥75% primary):806.09 and 806.04 — both are "not elsewhere classified" (NEC), mostly affecting 40s–50s

This suggests that central/anterior cord syndromes, although rare, are almost always serious enough to be the main reason for ICU admission. Also, injuries in younger adults (20s–50s) often present with more severe patterns.

---

5. Machine Learning Models

Multiple models were used to predict ICU LOS:

1. **Linear Regression:** A baseline model, incorporating features like age, ventilator duration, and ED duration.
2. **Neural Networks:** This advanced model integrates both structured and unstructured data for predicting ICU LOS.

Neural Network Model for Predicting ICU Length of Stay

A neural network model was developed to predict ICU length of stay (LOS_ICU_DAYS) for patients with spinal cord injury. Because the raw LOS values were highly skewed, we applied a log transformation (i.e., using log1p) to create the target variable LOS_ICU_LOG. This transformation helped normalize the distribution and reduce the impact of extreme values, leading to improved model stability and performance.

Data Preparation:

- The cleaned dataset was used as the starting point.
- The categorical variable INSURANCE was one-hot encoded, and LOS_ADM_DAYS (length of hospital stay) was retained as a numerical predictor.
- The final feature set comprised LOS_ADM_DAYS along with the dummy variables for INSURANCE.
- The dataset was then randomly split into training and testing sets in an 80:20 ratio.
- Feature values were standardized using StandardScaler to help the neural network converge more reliably.

Neural Network Architecture:

- We built a Sequential neural network with three layers:
    - **Input Layer:** A Dense layer with 16 neurons and ReLU activation, receiving the standardized input features.
    - **Hidden Layer:** A Dense layer with 8 neurons using ReLU activation.
    - **Output Layer:** A Dense layer with 1 neuron and linear activation, which provides the predicted value of LOS_ICU_LOG.

Model Evaluation:

- The Linear Regression model we developed to predict ICU length of stay (LOS_ICU_DAYS) for patients with acute spinal cord injuries. To address the skewness of the LOS distribution and meet the assumptions of linear regression, we applied a logarithmic transformation to the target variable. The final model included a range of one-hot encoded features representing admission type, insurance status, diagnoses, and other patient characteristics. The model achieved an R^2 of approximately 0.486, meaning it explained around 48.6% of the variance in log-transformed ICU LOS. Key predictors included admission type and insurance—patients admitted electively generally had shorter ICU stays, while those with certain insurance types, such as Medicaid, tended to stay longer. The hospital mortality flag (HOSPITAL_EXPIRE_FLAG) was also associated with increased LOS, indicating that patients who did not survive their hospital stay typically spent more time in the ICU. Some diagnosis categories and discharge destinations also showed meaningful effects. However, the model revealed signs of multicollinearity, likely due to the large number of dummy variables, which may affect the reliability of individual coefficient estimates. Despite this, the model provided valuable insights into how administrative and clinical factors relate to ICU resource utilization for this patient population.
- Neural Network Model achieved an R^2 of -0.08, showing a lot of variability that remains unexplained by the current features and architecture. The neural network is training smoothly on the log-transformed ICU LOS target but can only explain ~8% of the variance in the test set given the limited features and sample size. It's a reasonable baseline, but more data, more features, or model tuning would likely improve the predictions on real ICU length of stay. MSE =140.32: On average, the squared error is ~140 days^2. MAE = 7.26: The average absolute error is ~8.26 days. So, our model's predictions are off by about 8 days on average. This is due to the extreme outliers (like 54.17, 59.37, 63.37 days) which bring skewness to the dataset and the model.

---

6. Results and Discussion

Key Findings:

1. **Model Evaluation**:
    - The **Linear Regression** model was the primary model used to predict ICU Length of Stay (LOS). The **OLS Regression** results indicated that **LOS_ADM_DAYS** was a significant predictor of ICU LOS, with a **coefficient** of **0.0547** and a **p-value** of **0.028**, suggesting its strong impact on ICU length of stay.
    - The **R-squared value** for the **Linear Regression** model was **0.472**, meaning that **47.2%** of the variance in ICU LOS can be explained by the model. This indicates a moderate fit and provides a useful, interpretable model for predicting ICU LOS.
    - The **Mean Absolute Error (MAE)** of **5.4 days** suggests that, on average, the model's predictions deviate from the actual ICU LOS by **5.4 days**. Although this is a reasonable estimate, there is room for improvement in predictive accuracy.
    - **OLS Regression** results show that **INSURANCE_TYPE** (such as Medicaid and Medicare) had less impact on ICU LOS, as indicated by the high **p-values**, suggesting that insurance types are not strong predictors of ICU LOS in this context.
    - The **Neural Network Model** indicates no major signs of overfitting: Training and validation losses converge nicely, so the network is generalizing about as well as it can with these features.

2. **Significant Predictors**:
    - The most significant predictors for ICU LOS included **LOS_ADM_DAYS**, **ED Duration**, and **ICD9 codes for cervical spinal cord injury**.
    - **Age** also showed a moderate impact, with older patients generally requiring longer ICU stays.
    - Other clinical features, such as **ventilator duration** and **admission type**, were expected to be important, but based on the models, these variables might require further analysis to explore their full potential in predicting ICU LOS.
    - Neural networks are often regarded as "black box" models, making it challenging to extract direct measures of feature importance. However, through sensitivity analyses and ablation experiments, LOS_ADM_DAYS emerged as the most influential predictor, which is clinically intuitive: patients with longer overall hospital stays tend to experience longer ICU stays. In contrast, the insurance-related variables did not exhibit a strong independent effect in the current model configuration. This finding implies that while socioeconomic factors might have some bearing in certain contexts, the immediate length of the hospital stay (captured by LOS_ADM_DAYS) is more directly associated with the duration of intensive care. These insights not only validate the model's partial predictive capability but also underscore the need to explore additional clinical predictors to capture the variability of ICU resource use more comprehensively.
3. **Key Observations**:

- The **ICD9 Code** associated with **cervical spinal cord injury** (**806.0–806.09**) played a prominent role in understanding ICU LOS, as **SCI patients** required longer ICU stays due to complications and the nature of their injuries.
- **Emergency admissions** consistently had longer ICU stays compared to **elective admissions**, highlighting the severity of trauma cases and their impact on resource utilization.

---

7. Conclusion

This study successfully demonstrated that **Linear Regression** can be used to predict **ICU Length of Stay (LOS)** for **acute traumatic spinal cord injury (SCI)** patients, providing valuable insights for hospital resource allocation. Despite using a **simple linear model**, the results indicated that significant predictors, such as **LOS_ADM_DAYS**, **age**, and **ICD9 codes** for cervical SCI, are crucial for understanding ICU LOS.

While the **R^2** value of **0.472** indicates that the model captures nearly half of the variance in ICU LOS, there is still potential for improving the accuracy of predictions. The study shows that more advanced models (such as neural networks or ensemble methods) could further enhance predictive performance.

The findings suggest that ICU LOS for **SCI patients** is influenced by multiple factors, including the nature of the injury, emergency vs. elective admissions, and patient demographics. By identifying these predictors, this study offers a foundation for future research and the development of more accurate predictive tools in ICU management.

Limitations:

Model Generalizability:The model was trained and tested on data from a single hospital's ICU admissions. As a result, the model may not generalize well to other hospitals or healthcare settings, especially those with different patient demographics, treatment protocols, or healthcare infrastructure.

Data Quality:While the **MIMIC-III** dataset is comprehensive, certain records contained missing values or inconsistent data. This could have affected the performance of the model, particularly for features like **ICD9 codes**, **ventilator duration**, and **lab results**.

Simplification of Features:The **OLS Regression** model used in this study only considered a limited set of features. More complex models could integrate additional factors, such as real-time monitoring data or multidimensional treatment protocols, to better capture the complexity of ICU LOS predictions.

Future Work and Possible Next Steps:

- **Add More Relevant Clinical Features**: Vitals, comorbidities, lab results, or severity scores.
- **Expand Dataset**: 123 training samples is small for a neural net.
- **Try Different Model Types**: Random forest or gradient boosting sometimes handle tabular data better.
- **Hyperparameter Tuning**: Adjust layer sizes, learning rate, or dropout.
- **Cross-Validation**: For small data, cross-validation can give a better estimate of out-of-sample performance.

External Dataset Validation:To improve the generalizability of the model, additional datasets from multiple hospitals or healthcare systems should be incorporated. This will help validate the model's predictions and ensure its robustness across different patient populations and clinical environments.

Incorporation of Real-Time Data:Future models could incorporate **real-time data** from ICU monitoring systems to provide dynamic, up-to-date predictions of ICU LOS. This could improve the timeliness and accuracy of ICU resource allocation in hospitals.

Feature Expansion:**Additional features**, such as **clinical notes** (using **NLP** techniques) or more granular data on patient treatment and rehabilitation, could further enhance the predictive power of the model.

Survival Analysis:**Survival analysis** techniques, such as **Cox Proportional Hazards**, could be explored to predict time-to-event data for ICU discharge, providing an alternative perspective on ICU LOS prediction.

---

8. Appendices

Appendix A: ICD-9 Codes for Cervical Spinal Cord Injury

The **ICD-9** codes used to identify cervical spinal cord injuries (SCI) in the **MIMIC-III** dataset are as follows:

806.0 - 806.09: Cervical Spinal Cord Injury (SCI)

- **806.0**: C1–C4 fracture with spinal cord injury, unspecified
- **806.01**: C1–C4 fracture with complete spinal cord injury
- **806.02**: C1–C4 fracture with incomplete spinal cord injury
- **806.03**: C1–C4 fracture with central cord syndrome
- **806.04**: C1–C4 fracture with anterior cord syndrome
- **806.05**: C5–C7 fracture with spinal cord injury, unspecified
- **806.06**: C5–C7 fracture with complete spinal cord injury
- **806.07**: C5–C7 fracture with incomplete spinal cord injury
- **806.08**: C5–C7 fracture with central cord syndrome
- **806.09**: C5–C7 fracture with anterior cord syndrome

These codes help in classifying the type and severity of cervical SCI, which are significant factors in predicting ICU LOS.

Appendix B: Python Code

The following Python code was used for data preprocessing, feature engineering, and model training. This section includes key scripts for the analysis:

- Data Preprocessing and Feature Engineering
- Model Training and Evaluation (Linear Regression)
- Visualization Scripts

Appendix C: Cross-Validation Results

The following performance metrics were evaluated for each model using **cross-validation**:

1. **Linear Regression Model**:
    - **R^2**: 0.472
    - **Mean Absolute Error (MAE)**: 5.4 days
    - **Root Mean Squared Error (RMSE)**: 7.5 days
2. **Neural Network Model** (baseline model):
    - **R^2**: 0.21
    - **Mean Squared Error (MSE)**: 111.21 (average squared error of ~111 days²)
    - **Mean Absolute Error (MAE)**: 7.3 days

The **Linear Regression** model performed better, explaining **47.2%** of the variance in ICU LOS, while the **Neural Network** model showed substantial variability and required further tuning.