Predicting Length of ICU Stay in People with Acute Traumatic Spinal Cord Injury

By: Stefan Jafry, Shrivar Naidu, Tony Ngo, Ethan Elias, Mustafa Mohiuddin

Problem:

Background:

"Intensive Care Units (ICUs) provide critical, life-saving care for patients with severe medical conditions. However, ICU stays are expensive (~3x the cost of a regular hospital bed) and require careful management of limited medical resources such as ventilators, beds, and staff. The ability to accurately predict ICU Length of Stay (LOS) is crucial for:

- Optimizing hospital resource allocation (bed availability, staffing, ventilators, etc.)
- Reducing costs for Hospitals
- Timely Discharge Planning

In 2017, a systematic review was conducted for 31 ICU LOS machine learning models, all performing poorly (R^2 of 0.05-0.28). Furthermore, an already existing model called APACHE IV attempted to predict ICU LOS for Spinal cord injury, performing poorly, especially for specific groups. APACHE IV achieved an R^2 of only 0.21. These models leave significant room for improvement."

Overall Research Topic:

To create a machine learning model to predict ICU LOS in Acute Traumatic Spinal Cord injuries (SCI) patients. Our primary objective is to develop a model that outperforms APACHE IV (which achieved an R^2 of 0.21)

Subtopic:

To conduct a more thorough analysis, we will create more niche machine-learning models tailored to specific SCI subgroups. We will use unsupervised clustering to identify natural patient subgroups based on data patterns and use those subgroups as the basis for our more specific machine-learning model.
Example Approach:
If clustering reveals age as a significant factor, we will train separate machine learning models for:

1. Adults (18-60 years)
2. Older Adults (60+ years)

This approach may uncover hidden patterns—e.g., patients 60+ may be 30% more likely to require surgery, allowing hospitals to allocate resources like surgical beds and post-op care preemptively.

Expected Benefits of Niche Models:

Higher Prediction Accuracy: By training on Homogeneous subgroups, the models will capture more precise trends
Improved Clinical Interpretability**:** Helps doctors and hospitals make more accurate evidence-based decisions.

More Effective ICU Resource Management: Hospitals can anticipate bed turnover, surgery demands, and ventilator needs.
By creating a generalistic and more niche model, we offer the best of both worlds, allowing for broad applicability and more specific applicability and insight.

Impact:

Predicting LOS for ICU Spinal cord injury patients significantly impacts hospitals, patient care, hospital efficiency, cost reduction, and many other factors. Predicting LOS can help create personalized treatment and risk reduction early. For example, if the model flags a high-risk patient early, doctors are able to implement early treatment leading to lower mortality rates, faster recoveries, and shorter ICU stays. Furthermore, it can reduce costs for patients and hospitals, avoiding unnecessary hospital days.

Impact Of Overall Research Topic:

Pros:

- A more general machine learning model has broad applications and can predict LOS for any SCI patient, regardless of age, injury, or severity.
- It identifies broader patterns and trends that could apply to other hospitals, not just this specific institution.
- These models are more easy to develop and require less maintenance.
- It also serves as a resource that institutions can use to estimate overall ICU demand for SCI patients, optimizing bed availability and ventilator allocation.

Cons:

- A more general model will have lower prediction accuracy, especially for subgroups needing more nuanced care. ‘A one-size-fits-all' model may fail to capture important differences between SCI patients.
- Less interpretability and trustworthiness for clinicians, especially when clinicians need condition-specific reasons.
- It can misallocate resources, especially in more niche cases.

Impact Of Subtopic:

Pros:

- More homogenous data leads to higher prediction accuracy, especially among subgroups.
- It can provide more relevant insight for physicians in improving clinical interpretability and decision-making.
- More accurate for resource allocations, especially for more niche cases

Cons:

- Reducing scope can lead to small sample sizes. If we over-segment, the dataset will become too fragmented, reducing effectiveness. In some cases, the data may be too small to create an effective model.
- Overfitting leads to less generalizability across ICU settings; for example, a model performing exceptionally in one hospital may fail in another.
- It is more computationally expensive and requires more maintenance.

Impact Of Using Both:

Creating general and niche models will help mitigate each model's cons while maintaining its pros. It combines broad applicability with high accuracy for specific patient
groups. Using unsupervised clustering to create subgroups will help create accurate and more impactful niche models.

Cons:

- Increased computational complexity and maintenance.
- It could potentially require more training data and validation (unsure as of now).
- Potential conflicting conclusions in models (create a rule-based system: IF: patient fits niche category, use niche model, ELSE: use general model) can be a potential solution.

Data:

Tell us how large your dataset is:

We are only using the ICU and HOSP models from MIMIC-IV, which are significantly smaller than the full version.

ICU Dataset: This model includes data related to ICU patients, such as vitals and medication.

key Tables & Approximate Row Counts:

- IcuStays ~70,000 rows (one row per ICU stay)
- ChartEvents ~300 million rows (largest table, includes vitals and bedside data)
- IngredientEvents ~50 million rows
- InputEvents ~10 million rows
- OutputEvents ~5 million rows
- Procedure Events ~2 million rows
- DatetimeEvents ~1 million rows
- D_Items~40,000 rows

Hosp Data Set: This module includes patient data hospital-wide, such as patient age, diagnosis, and lab results, among many other tables. Since not all data in this set is relevant to our study, we decided on only choosing admissions, patients, diagnoses, prescriptions, HcpsEvents, labevents, microbiology events, OMR, services, and transfers:

Key Tables & Approximate Row Counts:

- Admissions ~200,000 rows (one per admission)
- Patients ~200,000 rows (one per patient)
- Diagnoses_Icd ~2 million rows (ICD codes for each admission)
- Prescriptions ~10 million rows (medication orders)
- HcpsEvents ~5 million rows
- LabEvents ~50 million rows
- MicroBiologyEvents ~5 million rows
- OMR ~1 million rows
- Services ~1 million rows
- Transfers ~3 million rows

The estimated size of all the data we choose is approximately 80-100GB. It is important to note that this data includes patients from all demographics and injury types. Since our study only focuses on ICU SCI patients, SQL will be used to remove all irrelevant non-SCI patients, which will reduce the size by a substantial amount.

How was the data collected? Is there any part that affects your analysis?:

- The MIMIC datasets originate from Beth Israel Deaconess Hospital in Boston. Data was collected during routine care, such as vital signs, laboratory results, medications, and treatment. The information was put into the hospital's electronic health record, where the patients' identification was removed to protect privacy. It was then put into MIMIC databases.
- Since it was mainly collected from a single hospital, our analysis might not be reliable for the general population as the training specifically came from one source. The range of this dataset in terms of time is extensive enough to the point where medical practices and ICU protocols might have changed, so our analysis could be less impactful as time goes on if practices continue to evolve. Removing patient identification could leave out important information relevant to our analysis and for a more nuanced predictive model.
- Inaccurate data entry could shift our analysis if the data is inputted incorrectly throughout the datasets.

Comment on the cleanliness of the data:

- Some check in and leave times are missing, which could negatively influence our results to produce inaccurate conclusions, meaning that we will have to start by cleaning these values
- in d_hcpcs.csv, long_description has an entire column with NaN values. We plan to remove it due to its redundancy and use only short_description. A few of these rows also had 'invalid code' as their short description, so we plan to do more data cleaning to eliminate those error values.
- In labevents.csv, many comments had NaN values. We believe this was because no further comments were available; however, the data set may have errors, which will be checked further.
- For prescriptions.csv, the form_rx column has only NaN values.
- Since much of the data is relevant to patients without SCI, we will need to transform the data and create new tables that only include SCI patients for our analysis.

Are there any variables not present in the dataset that would be nice to have?:

- spinal_injury_location would provide more readability and classification in our dataset.
- spinal_injury_type is also helpful in deducing the exact type of spinal injury, such as anterior cord syndrome.
- cause_of_injury is another variable that would tell us the severity of the injury and be a significant tool to help us predict the stay time
- therapy_response would be a useful variable to have as it would describe how the patient responded to treatment and therapy. This would strengthen our model as it was excluded in APACHE IV and has been proven to impact the length of stay.
- Having a table related to rehabilitation readiness and more recovery metrics, especially related to ICU SCI, would greatly benefit our study. For example, physical therapy, muscle strength scores, cognition, and depression screening data would help deduce what rehabilitation tactics lead to faster recoveries. Sadly, this data is either not documented at all or not done properly in the MIMIC-IV database.

Methods:

Variables:

Our scope across all CSVs' will be narrowed to variables with values that describe or are related to acute spinal injury. As of now, we have decided to focus on variables that relate to such, as well as a potential impact on the length of stay. The correlation between these variables and length of stay will be found through the process of unsupervised clustering, and the results of this process will be validated by computing the correlation matrix of length of stay versus all other relevant variables.

For all ICU csv's:

chartevents.csv

Relevant variables in chartevents.csv are:

- subject_id (id of the patient)
- hadm_id(hospital admission id)
- stay_id (which icu the patient was in)
- itemid (tells us the code for the measurement recorded)
- value (how much the measurement is)
- valuenum (numerical representation of the value if applicable)
- valueuom (the unit of measurement of the value)
- Warning (issues with the data)

icustays.csv

The entirety of icustays.csv is very important since it has dense detail on admission time and whether the patient was transferred to another ICU. We will focus specifically on rows with the values Surgical ICU and Medical ICU, as they are most relevant to spinal injury.

ingredientevents.csv

Ingredientevents.csv looks at the medicine given to patients in the ICU. The variables we will analyze are:

- starttime(when the medicine began to administer)
- endtime(when medicine administration ceased)
- itemid (tells us the code of the medicine used)
- amount (the total amount of medicine administered)
- amountuom (amount unit of measurement)
- rate (the current rate of administration of the medicine)
- rateuom (rate unit of measurement)
- originalrate (the rate prescribed by specialists)

The reason for using this file is that certain medications given to patients could have a positive or negative effect on their recovery, altering their length of stay.

inputevents, outputevents & procedureevents

The CSVs for input, output, and procedure events share common columns; however, each CSV documents a different aspect of ICU treatment. Input events document all medicinal, fluid, or other treatments administered to the patient; output events document all fluids/substances removed

from/excreted by the patient; and procedure events note all medical procedures the patient underwent during their time in the ICU.
From these three csvs, our main focus would be on:

- itemid (procedureevents & outputevents)
- ordercategoryname(procedureevents)
- patientweight(procedureevents & inputevents)
- isopenbag (procedureevents & inputevents)
- continueinnextdept(procedureevents & inputevents)
- location & locationcategory (procedureevents)
- starttime & endttime (inputevents)
- amount & rate (inputevents)
- ordercatgeoryname & secondaryordercategoryname (inputevents)
- ordercomponenttypedescription (inputevents)
- totalamount, originalamount & originalrate (inputevents)
- value (outputevents)

Our goal here is to find which specific events are related to spinal injury and the factors involved in these events that could play a part in making a patient stay in ICU for longer.

For all Hosp csvs:

admissions.csv & diagnoses_icd.csv

We will also use the entirety of these CSVs in our investigation. Unlike Apache IV, we will also consider deathtime, which is present in admissions.csv.

hcpcsevents.csv

Our focus for this file will be on short_description and hcpcs_cd, as these will tell us what medical services the patient received. We will link the ones related to spinal injury and length of stay to our model. 'Hospital observation per hour' was a value that caught our attention and will be investigated further.

labevents.csv

Lab tests and results may extend the length of stay depending on the severity of the results, so we will focus on the following:

- flag (whether the lab result is critical or not)
- priority
- comments (further comments on lab results)
- value & valuenum (the lab results)
- valueuom (unit of measurement for lab results)

microbiologyevents.csv

One of the significant shortcomings of Apache IV was that it did not include complications in its prediction algorithm. Here, we will observe the effect of other complications in addition to spinal injury by looking at:

- test_name (what complications the test is checking for)
- interpretation (of test results)
- comments (of test results)

omr.csv

It contains important information on the patient's weight, height, BMI, and blood pressure. We will mainly consider the following:

- result_name
- result_value
- subject_id

patients.csv

We will want to consider:

- Gender (men may be more/less likely to recover from spinal injury, so longer ICU staying period)
- anchor_age, anchor_year, anchor_year_group (refer to the time variable at the time the data was recorded, which could be important in predicting length of stay)
- dod (refers to the date of death)

prescriptions.csv

Certain prescriptions could affect the length of stay. We will look at:

- starttime
- stoptime
- prod_strength (how much of the drug per pill)
- drug
- drug_type

services.csv

prev_services and curr_services may be important to our investigation as they dictate a shift in operations performed on the patient, i.e., a patient may go to ORTHO and NSURG(neurosurgery) for spinal complications, which can increase their length of stay in ICU and recovery.

transfers.csv

Here we shall investigate:

- eventtype(possible values are: 'ED,' 'discharge,' 'admit')
- careunit
- intime
- outtime

These variables will be needed to determine how long patients stay in an ICU, in which unit, and for what purpose.

Will you need to create new variables to conduct your analysis? What new variables are they?:

- We must create a Length of Stay variable as it is not explicitly stated in our data set. This will be done by finding the difference between 'admittime' and 'dischtime' in the 'admissions.gz.csv' file. We will also have to include 'intime' and 'outtime' from 'transfers.csv'. 'dod' and 'deathtime' should also be included in our calculation, as if present, they are equivalent to length of stay.
- Since many csvs share the same types of variables but for different purposes, such as 'value' and 'amount', we will have to reformat column names to specifically state what variable for what purpose is being accounted for.

- To further broaden our dataset for low variability in our model, we will include spinal_injury_loc, spinal_loc_des, and spinal_injury_type. spinal_injury_loc have the possible values C1, C2, C3, C4, C5, C6, C7. C1 to C4 (upper cervical spine), and C5 to C7 (lower cervical spine). spinal_loc_des will specify 'upper cervical spine' or 'lower cervical spine,' and spinal_injury_type will specify the exact type of injury, such as 'central cord syndrome,' and would map to corresponding injury codes that currently exist in the dataset.

Visualizations:

To analyze the relationships between variables in the study on predicting ICU stay length for individuals with acute traumatic spinal cord injury, a variety of visualizations and statistical techniques will be utilized.

- A **histogram** and **boxplot** will help examine the distribution of ICU stay length, identifying patterns such as skewness and potential outliers.
- **Scatter plots** will explore the relationships between ICU stay and key numerical variables like injury severity, age, and comorbidities, revealing potential correlations.
- A **correlation matrix heatmap** will highlight how different numerical variables relate to one another, while bar charts will compare ICU stay lengths across categorical factors such as injury type or the level of paralysis.
- **ROC curves** will also be beneficial for determining how accurate our model is.
- As more in-depth analysis is needed, regression models such as multiple **linear regression** or machine learning approaches will help quantify the influence of various predictors on ICU stay.

Example Of Potential Plots:

Comparison Type	Potential Plot Type	Example Use Case
Continuous vs. Continuous	Scatter plot	HR vs. BP in SCI patients
Continuous vs. Categorical	Boxplot / Violin Plot	Blood pressure across ICU units
Categorical vs. Categorical	Stacked Bar Chart	Infection types by ICU admission source
Time Trends	Line Chart	Lab test values over ICU stay
Multivariate	Pairplot / Heatmap	Correlation between vitals

Pick one of these visualizations and explain how it will help you solve the problem/answer the research question:

- **A scatter plot with a regression line** is a valuable tool for understanding how injury severity impacts ICU length of stay (LOS) in patients with acute traumatic spinal cord injuries (SCI). By plotting individual patient data, this visualization helps identify whether there is a clear pattern. For example, do more severe injuries consistently lead to longer ICU stays? If the points form an upward trend, it suggests that injury severity is a strong predictor, which could help hospitals better plan for things like bed availability, staffing, and ventilator use. On the other hand, if the data points are scattered without a clear pattern, it would indicate that other

factors—such as age, pre-existing conditions, or need for surgery—play a more significant role in determining ICU stay.

- This scatter plot is especially useful as a first step in building machine-learning models for predicting ICU LOS. The model can place more weight on injury severity if a strong relationship exists. However, if the relationship is weak, it reinforces the need to create more specialized models—for example, one model for younger patients and another for older adults. The current APACHE IV model, which attempted to predict ICU stay for SCI patients, struggled because it did not account for these differences and only achieved an R^2 of 0.21. By identifying distinct patterns through visualization, we can build more targeted machine-learning models that improve accuracy and provide better insights for doctors and hospitals.

Models and statistical tools:

List possible machine learning models or statistical tools that might be used to solve your problem:

When predicting the length of stay for patients with spinal cord injuries, we need to account for many factors, such as injury severity, the location of the spinal injury, and prescriptions. Considering that LOS is a continuous outcome, several potential models can predict LOS for ICU SCI patients.

- **Linear Regression and multiple linear Regression (MLR)**—This model helps us understand how different factors (injury severity, age, comorbidities, etc.) impact ICU stays.
- **Random Forest Regression—**This model creates multiple decision trees and averages their results. It works well with categorical and continuous features and is robust to missing data.
- **Gradient Boosting Models** (XGBoost, LightGBM, CatBoost)—These models refine predictions by learning from past mistakes, improving accuracy over time. They can handle complex relationships like the relationships between vitals, interventions, and LOS and work well on large data like MIMIC IV.
- **Neural Networks** (Deep Learning) - This captures complex patterns in ICU LOS, especially if there are hidden interactions between factors.
- **Survival Analysis** (Kaplan-Meier Curves, Cox Proportional Hazards Model) - This is useful for modeling time-to-event data, helping us understand how different factors influence how long a patient stays in the ICU.
- **Clustering Techniques** (K-Means, DBSCAN, Hierarchical Clustering) - This helps identify natural subgroups of patients, such as younger vs. older patients or mild vs. severe injuries.
- **Causal Inference** (Propensity Score Matching, Instrumental Variable Analysis, Difference-in-Differences) - These methods help determine whether specific treatments or patient characteristics cause longer ICU stays.

Comment on the challenges and the limitations of your proposed models or statistical tools:

- **Linear Regression Models** may prove inaccurate if the data follows a non-linear trend and may not accurately predict the length of stay.
- **Random Forest Regression** requires a lot of computational power, especially for a large dataset like the one we are working with.
- **Gradient-boosting Models** are more sensitive to outliers, meaning that outliers may have more residuals. Because the next tree uses these outliers, it will consequently be less accurate.

- **Neural Networks** are 'Black box,' meaning that we won't know exactly why we got a specific result for the length of stay.
- **Survival Analysis Models** use specific assumptions; if a data point does not meet them, the output for the length of stay can be unpredictable.
- **Clustering** may segment our dataset too much, creating overfitting.
- **Causal Inference** through Propensity Score Modelling can create bias in our dataset if irrelevant variables are included or if the data's relationship is incorrectly specified, i.e., as non-linear when it is actually linear.

Describe which methodology you will follow to train or validate your models or statistical tools:

1. **Data Cleansing, Transformation, and Feature generation using SQL:**

We would gather all relevant data from our dataset and SQL queries to filter out incomplete data and filter in data related to spinal cord injury and length of stay. Then, we would aggregate our data and handle any irrelevant information. We would also create the aforementioned required variables to broaden our data set.

2. **Creating a general LOS model:**

This would involve using a linear or logistic regression model (depending on the r-squared we get) and plugging in all our related variables into the Regression. Our test-training split would also occur here.

3. **Unsupervised clustering:**

Here, we will segment the data into similar groups, such as similar arrival and departure times. We shall then create two new models depending on how the data was clustered according to each cluster's relevance to ICU stay and spinal injury.

4. **Comparison of models to APACHE IV:**

Visualizations will be made to compare all of our models, including ROC curves and precision-recall curves. We will also use other performance metrics, such as R-squared and Root Mean Squared Error, for our comparison.

Concerns:

The MIMIC IV data set greatly improved from the MIMIC III dataset. It has a better structure, more streamlined data, updated coding standards, and more data, which increase representations, among many other improvements. Despite all this, there are concerns regarding data quality, lack of generalizability, technical challenges, the difficulty of comparisons/evolving medical practices, and overall concerns about our machine models.

Data Quality:

Missing Data: Data can sometimes be missing or incomplete, potentially causing important information to be lost. Data imputation, removing NULL values/rows/columns, will be necessary to conduct a thorough analysis. For example, some admission and discharge times are missing, reducing data quality.

Data Masking: Although an important step to ensure patient privacy, specific values were tampered with, which can potentially skew data. For example, all patients 89+ are listed as 91 years old, which can result in a loss of granularity, especially regarding elderly patients.

Generalizability Issues:

MIMIC-IV captures data on one hospital, the Beth Israel Deaconess Medical Center (BIDMC), in Boston, Massachusetts, USA. This means our findings may not directly apply to different hospitals. If our models perform poorly in other hospitals, they can only benefit one small demographic rather than a larger population.

Technical Challenges:

Complex data structure: Our data set consists of multiple relational tables that require SQL to join and query to create valuable data that our machine-learning models can interpret.
Difficulty accessing ICU SCI patients: MIMIC IV contains data on every patient admitted to BIDMC, meaning lots of data filtration, querying, and creating new tables will be required to find relevant data for our models.

Comparison Difficulty/Evolving medical practices:

Evolving Medical Practices: Clinical protocols, treatments, hospital guidelines, and changes in patient trends (ICU trends, evolution of medicine, mortality) are in constant flux, especially over time. This will affect the relevance of our past data, making our models irrelevant and, therefore, unusable. For example, if lumbar surgeries become more advanced in 10 years and thereby lead to faster recovery, our model will not incorporate this automatically and will become less accurate.
Comparison Difficulty: Different hospitals and countries have varying documentation standards, data, and trends, making direct comparisons challenging.

Machine Learning Models Concerns:

Complexity/Overfitting: MIMIC IV contains dozens of tables with data on subjects such as vital signs, prescriptions, and even the nurse assigned to the patient. Many of these features can have high correlation or redundancy, leading to overfitting, unnecessary noise, and potentially irrelevant features. Machine learning models may not translate to other institutions, making them less effective and less likely to learn generalizable patterns.
Variability: ICU patients undergo rapid physiological changes, with their lab results, vitals, and treatments changing hourly. If our model does not account for the time-series patterns, they may learn static relationships that do not hold over time.

Sources:

Southwest Journal of Pulmonary, Critical Care and Sleep - CRITICAL CARE - The Explained Variance and Discriminant Accuracy of APACHE IVa Severity Scoring in Specific Subgroups of ICU Patients

Critical Care Medicine

Validating the APACHE IV score in predicting length of stay in the intensive care unit among patients with sepsis | Scientific Reports

Which Models Can I Use to Predict Adult ICU Length of Stay? A Systematic Review - PubMed

Online ICD9/ICD9CM codes

Gradient Boost for Regression - Explained

4 Disadvantages of Neural Networks | Built In

Survival Analysis - What It Is, How It Works, Pros And Cons

Precision-Recall Curve | ML - GeeksforGeeks

Our background research can be found here:
https://docs.google.com/document/d/1aGFRkoSWa_1V8HdOurLEDI65c30Y2D3W1VJ4qmnbchE/edit?usp=sharing