



Data Science Mission - MIMIC IV

**Early Prediction of Prolonged ICU Stay at Second Admission:
Replicating and Expanding a Drug-Based Machine Learning Approach Using
MIMIC-IV**

MSc DSAIS

AUTHORS

Matilde Monti - Lou de Gaetano Nerino de Visconti - Sicheng Huang
Himanshu Midha - Joaquin Arias

Contents

1	Introduction	1
2	Literature Review	2
3	Data	5
3.1	Pipeline Description	6
3.2	Data Extraction	7
3.2.1	Dataset Extraction (Reference Study)	7
3.2.2	Additional Clinical Data Integration	8
4	Methods	9
4.1	Modeling	9
4.2	Discharge Note Embedding	10
4.3	Model Hyper-Parameter Tuning and Calibration	11
5	Results	13
5.1	Data Exploration	13
5.2	Results from Paper	13
5.3	Results from Paper extracted by us	15
5.4	Model results from an additional dataset with discharge notes	18
5.5	Feature Importance	19
5.5.1	Feature Importance on discharge notes dataset	22
6	Discussion	23
7	Implementation (Streamlit)	24
8	Conclusion	28
9	Future Work	29
	List of Figures	I
	List of Tables	II
	References	III
	Appendix: Exploratory Data Analysis	VI
.1	Univariate Analysis	VIII
.1.1	Age Distribution	VIII
.1.2	Length of 1st Admission distribution	X

.1.3	Length of 2nd Admission distribution	XI
.1.4	Disease Prevalence	XII
.2	Bivariate Analysis	XIII
.2.1	Age V/s Gender	XIII
.2.2	Age V/s Disease Correlation	XIV
.2.3	Disease V/s Length of 1st admission	XV
.2.4	Top 10 Correaltions between Age, Length V/s Medications	XVI
.2.5	Top 10 Correlated Disease Pairs	XVII

1 Introduction

Accurately predicting patient outcomes in the intensive care unit (ICU) is a central challenge in critical care and a subject of growing investigation in applied data science. The ICU is one of the highest-cost and highest-demand areas of a hospital. Beds are limited in number, personnel are specialized, and treatment is expensive. Of all of the various patient outcomes that clinicians and administrators would like to forecast, ICU length of stay (LOS) is of particular interest. LOS is defined as the time interval between the hospital admission and discharge during a given admission. Long ICU lengths of stay are associated with higher complication rates, higher mortality rates, and higher costs of care. They also create ICU congestion that can slow the treatment of other patients and impact system-wide efficiency. Identifying which patients are going to require extended ICU treatment enables better planning of resource utilization, optimization of bed turnover, staffing needs, and more efficient care planning earlier in the clinical course.

Despite the clear utility of ICU LOS prediction, current machine learning (ML) techniques are typically narrow in scope. Most models are intended to generate outcome predictions based on observations made during a single ICU stay and on static features such as demographic and admission-time laboratory tests. While useful, these techniques do not capture the dynamic nature of a patient's clinical course or account for how a patient's prior ICU course might influence future outcomes.

The goal of this study is to build a machine learning predictor that can predict whether a patient's future ICU stay will be prolonged (≥ 3 days). Our approach integrates information from two key sources: (1) patient history on the first ICU stay and (2) contemporaneous clinical data, medications, labs, and vital signs.

In developing our modeling pipeline, we built on the methodology outlined in a recent paper by Kuo et al. (2024) [1]. The paper introduced a formal machine-learning framework for LOS prediction on the MIMIC-IV database using a combination of demographic information, comorbidities, and vital

signs. While their work was done using a prior iteration of MIMIC-IV, we applied the data preprocessing pipeline used in this study. Their preferred predictive models, primarily tree-based learners such as XGBoost, LightGBM, and Random Forests, served as our baseline performances. From there, we made a series of improvements that included revised cohort selection criteria, more advanced methods of feature selection, and hyperparameter tuning to improve the discrimination and generalizability of models.

The rest of this paper is organized as follows: Section 2 provides a review of the related literature on ICU LOS prediction, highlighting areas that our study aims to bridge. Section 3 provides an overview of the dataset and the steps followed to prepare the final data used in the analysis. Section 4 presents the learning models leveraged, as well as the evaluation metrics employed and optimization methods adopted. Section 5 discusses the results, comparing them with the baseline research, while section 6 presents our findings and insights. By developing a reproducible, interpretable, and clinically relevant machine learning model, this study advances predictive analytics in critical care. Specifically, it demonstrates that incorporating a patient's ICU history into real-time prediction pipelines can enhance risk stratification and facilitate faster and better-informed clinical decision-making. From an operational viewpoint, better LOS prediction can facilitate ICU capacity management by hospitals, reduce delays in care, and ultimately improve patient outcomes and resource utilization.

2 Literature Review

Machine learning predicting models in critical care have evolved in the last years, especially concerning patient outcomes (mortality, readmission risk, ICU LOS). These models may be a great help in hospital resource optimization and patient risk stratification. MIMIC-IV, one of the largest datasets in the medical field, is powering this research. It is a large, de-identified critical care database developed by the MIT Laboratory for Computational Physiology. Its design accommodates longitudi-

nal and time-aware modeling, making it particularly well-suited for complex prediction problems like early detection of extended ICU admissions.

Zhang and Kuo (2024) [1] focused on the early prediction of ICU length of stay ($\text{LOS} \geq 3$ days or ≤ 3 days) occurring at the second ICU admission, based only on the data of the first ICU admission, specifically demographic, diagnosis, and medication-related features. Their approach was particularly unique as it included drug dosage information, specifically the average and sum values for commonly prescribed medications, which turned out to be important predictive signals. They evaluated five classifiers (Logistic Regression, Random Forest, SVM, AdaBoost, and XGBoost) among which the Random Forest gave the highest AUC (0.716) as well as the best calibration of the classifier [1]. Their study confirmed that it is indeed possible to estimate a patient's risk of prolonged ICU LOS within the first few days of the second admission given structured, first-admission data.

Beyond this baseline, we explored the wider literature around predicting ICU LOS specifically with the MIMIC datasets. Syed et al. compiled a systematic review of ML applications in ICU, leveraged MIMIC-III and IV [2], demonstrating that these data sets are quite heavily adopted for ML implementations, regardless of which ML model was leveraged (random forests, SVMs, or various deep learning architectures). Their results highlighted that static measurements such as lab tests, vitals, and diagnosis codes are the features that have higher predictive power, though challenges surrounding missing data, redundancy of features, and lack of standardization in preprocessing remain prevalent across studies.

To handle preprocessing issues, Gupta et al. (2022) [3] developed a comprehensive data processing pipeline for MIMIC-IV that influenced many of our own design choices. Their pipeline offered systematic approaches for high-dimensional and sparse data management, feature selection, and well-defined cohort inclusion criteria. This impacted our treatment of missing values, normalization of medication doses, and standardized diagnostic code groupings in our dataset [3].

Comorbidities have also been well-documented to predict outcomes. Sharma et al. (2022) investigated comorbidity transformations based on diagnosis codes extracted from the MIMIC-III database and discovered that using clinically informed grouping (for example, the Elixhauser comorbidity index) enhances performance and interpretability. In our project, we implemented a similar strategy, mapping ICD codes from the `diagnoses_icd` table into Elixhauser categories, yielding binary comorbidity indicators for each patient [4].

Furthermore, the performance of ML classifiers for ICU outcomes was compared in several studies. Cheng et al. (2022) tested several traditional ML models such as ANN, decision trees, and KNN, for cost-efficient ICU forecast tasks. They found that while deep learning approaches consistently outperform others, the improvements are marginal and hinged on computation costs and data preprocessing bottlenecks, a key aspect with an impact on reproducibility and clinical deployment [5]. We looked at tree-based and linear models as this was consistent with Zhang and Kuo and we prioritized interpretability and reproducibility.

Although previous LOS prediction studies rarely use medication data, its association with LOS is clinically relevant. Zhang and Kuo (2024) demonstrated that the addition of medication dosage-based features meaningfully improves model performance. The feature importance analysis also shows the importance of common medications like Phytonadione and Metoprolol Succinate XL [1]. Our project built directly from this insight, deriving medication features from the second ICU stay as well as the first admission (as in the original study), which allowed our model to include more recent treatment indicators, while still preserving feasibility for early prediction.

A big discussion topic relating to MIMIC is fairness and generalizability. A 2024 preprint by Li et al. assessed the fairness of MIMIC-IV-derived models in sub-populations and documented performance discrepancies regarding age and ethnicity [6]. Our study did not include a fairness audit, which is a crucial area of further study as these models are deployed in real-life clinical environments.

Lastly, while we centered our project solely on using structured data, we also attempted to use natural language processing. The MIMIC-IV-ICD benchmark (2023) presented a multi-label classification task based on unstructured clinical notes and discharge summaries [7]. Although not directly connected to LOS, it also shows the potential of the with-text type of data in making predictive modeling complete and is an area we can extend our work on for future research.

Our work lies at the intersection of strong foundational work on ICU prediction and recent strides forward in feature design and clinical relevance. We closely adhered to the methodology of Zhang and Kuo (2024) as our baseline [1], then extended their work through additional features from newer MIMIC-IV data, re-tuning of classifiers, and incorporation of significant changes in patient state at second admission, all while preserving the early-prediction constraint of the original study. Our work adds to the literature by investigating the replicability of a previous high-impact predictive framework and its optimization given a novel dataset and extended clinical context.

3 Data

The data used for our study is MIMIC-IV (Medical Information Mart for Intensive Care IV) version 2.2, a large public electronic health record (EHR) database developed by the MIT Laboratory for Computational Physiology, in collaboration with Beth Israel Deaconess Medical Center (BIDMC). It contains de-identified clinical data for over 65,000 ICU patients admitted between 2008 and 2019. It is structured into multiple interrelated tables, classified into three general modules – core, hosp, and icu – that collectively offer extensive information on patient demographics, hospital admissions, as well as diagnoses and procedures, laboratory tests, vital signs, medication, and other time-stamped clinical events. All MIMIC-IV patients are assigned a unique patient ID to follow them longitudinally through multiple ICU admissions and hospitalizations. This structure is particularly well-suited to our study's objective.

For the study cohort definition, we selected patients with only two different ICU admissions, with the second being our target. The patient's data was further filtered following this criteria: patients who were given at least one frequently administered medication (defined as more than 1000 times of use across all patients) and diagnosed with at least one frequent diagnosis (defined as more than 10,000 times across all patients). To further improve the information in our dataset, different types of features were also extracted from each admission: vital signs, mortality rates, and discharge notes from the first admission.

The resulting dataset is structured for supervised learning: each row corresponds to the patients' first ICU admission, with labeled features and binary LOS outcome. This multi-source approach allows for learning both from the patient's prior risk-factor profile as well as their clinical course during the first admission, which we expect to yield more relevant and reliable estimates of prolonged ICU stay.

3.1 Pipeline Description

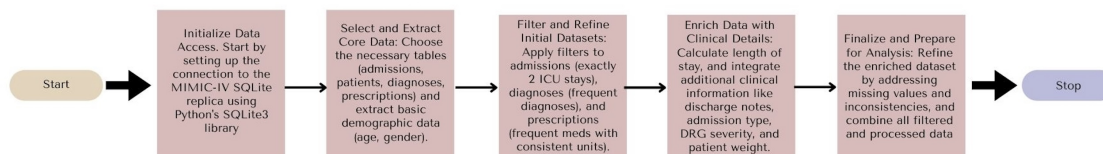


Figure 1: Data Extraction Pipeline

The dataset was constructed following the reproducible methodology delineated by Zhang and Kuo (2024) [1] by filtering and merging the data from the admissions, demographics, diagnoses, and prescriptions tables. To improve the quality of our data, additional information from the omr, drgcodes, and discharge tables was later incorporated to the dataset by merging on the patients' unique identifiers.

Prior to the dataset creation, the individual files were compiled into a complete database using the Python SQLite module, allowing easier handling of the data. Additionally, the relevant tables were indexed to facilitate data management. This structured approach ensures consistency and reproducibility, simplifying data filtering, merging, and testing during development.

3.2 Data Extraction

To generate our dataset, we did a structured extraction of information from the MIMIC-IV database. As described in our reference study [1], we utilized five data tables: admissions, patients, diagnoses_icd, d_icd_diagnoses, and prescriptions. A formal querying environment was established by connecting our local SQLite replica of the MIMIC-IV database to our Python environment using the SQLite3 module. This setup allowed us to directly execute SQL queries in Python, simplifying the filtering and merging of data across multiple tables while ensuring the pipeline's reproducibility and enabling systematic testing and updating during development.

After extraction and filtering, we created an intermediate dataset with the patients consistent with our inclusion criteria and further enriched it with demographic, diagnostic, and prescription-related characteristics from both ICU admissions. This dataset was used to form the foundation for the next step: machine learning-ready data cleaning and transformation.

3.2.1 Dataset Extraction (Reference Study)

The initial dataset was prepared following the methodology described in our reference study, replicating the data extraction and filtering procedures.

We first extracted patients' basic demographic information from the admissions and patients tables by filtering the data to include only patients with exactly two admissions. We then extracted diagnoses and prescription features from the respective tables as follows: for diagnoses, we first merged

`icd_diagnoses` with `d_icd_diagnoses` on both `icd_code` and `icd_version`¹. Then, to reduce the risk of overfitting, we filtered the patients who had at least one frequent diagnosis (defined as occurring at least 10,000 times across all patients).

A similar procedure was applied to prescriptions: the `prescriptions` table was filtered to only retrieve patients who had been administered medications with at least 1000 occurrences across the entire dataset. Additionally, to ensure data consistency, only patients who were prescribed medications with the same units were kept.

After applying these filtering criteria, we identified the 81 most common diagnoses and the 68 most-common medications following the baseline study's methodology. Our dataset comprised only 62 diagnoses compared to the 81 described by Zhang et al., resulting in 203 columns, instead of 220. Lastly, we calculated the length of stay for both the first and second admission by subtracting `admittime` from `dischtime`.

3.2.2 Additional Clinical Data Integration

To improve the accuracy of our model, we enhanced the information contained in the dataset by adding relevant patient information.

First, we added the discharge note for the first admission from the `discharge` table. Then, we added the `admission_type` from the `admissions` table [8], as well as the DRG severity and mortality scores. These two values indicate, respectively, the level of illness severity and the risk of mortality associated with the DRG classification² (values ranging from 1 to 4). Lastly, we added the patient weight, prioritizing the values from `inpuvents` and filling missing values with data from the `omr` table.

¹`icd_code` represents the diagnosis code assigned based on the International Classification of Diseases (ICD) system, while `icd_version` indicates the ICD revision under which the diagnosis code falls.

²The Diagnosis-Related Group (DRG) classification is a system used to categorize patients with similar clinical diagnoses.

Other vital signs were initially included, but due to a high proportion of missing values in these columns, we eventually decided to drop vital signs such as blood pressure and heart rate. The final dataset consists of 31,386 unique patients, with 216 features.

4 Methods

To guide our modeling strategy, we closely followed the methodology described by Zhang and Kuo (2024), who predicted the risk of a prolonged ICU length of stay ($\text{LOS} \geq 3$ days) at a patient's second ICU admission, using only structured data from the first ICU admission.

Their study showed that a small number of early features (Wang 2020; demographics, diagnosis codes, medication info, both cumulative and average doses) are predictive enough to give meaningful early warning, enabling targeted higher-risk patient mitigation efforts during admission.

Their framework was a methodological and experimental baseline for our project. We followed their path and built five supervised learning algorithms: Logistic Regression, Random Forest, Support Vector Machine, AdaBoost, and XGBoost. To improve the study, we also implemented a Naive Bayes algorithm, a model frequently used when analyzing medical classification problems [9].

4.1 Modeling

Before running the models, we encoded the categorical variable 'gender', assigning 0 to females and 1 to males, and scaled all non-binary variables.

After removing null values, duplicates, and patients with a length of first stay lower than 0³, we split the data into 50% training, 25% validation/calibration, and 25% test set. As the dataset was quite balanced (52.1% - 47.9%), the split was performed without stratification, as done by the reference

³A $\text{LOS} < 0$ indicates the death of the patient before reaching the hospital.

paper. This process was applied to both the original dataset retrieved from the reference study and the enhanced dataset with additional information and textual notes.

4.2 Discharge Note Embedding

To use the textual data in our predictive models, the discharge notes were processed to be included in a new structured dataset. The aim was to transform unstructured clinical text into a structured format, suitable for predictive modeling, with the aim to enhance the results previously obtained only with structured data.

Discharge notes are a comprehensive medical document created when a patient is discharged from the medical unit, summarizing the details of the patient's hospital stay. They serve for both patients and healthcare assistants, detailing the procedures and clinical results, as well as physicians' observations. Effectively processing them is a crucial step for extracting meaningful information to improve the predictive power and reliability of the models applied in subsequent steps. [10]

Before converting the medical notes into embeddings, it is necessary to preprocess the text. To do so, the stop words to be removed need to be defined: stop words are commonly-used words that can be seen as irrelevant for the purpose of text analysis, such as “the”, “is”, “in”, “on” and “at” [11]. As they appear frequently, they do not usually carry meaningful connotations about the topic being discussed. In the specific example of medical text, it is necessary to also include domain-specific stop words. For the study, we identified the following words: “patient”, “history”, “medications”, “symptoms”, “treatment”, “diagnoses”, “discharge”, “date”, “birth” and “sex”. Their removal reduces computational overhead and improves the efficiency of text processing. To ensure the quality of the cleaning process, negations were removed as stop-words, as in the medical field they convey important information regarding a patient's condition, procedures and test results. [12]

Therefore, the cleaning function follows the following process: firstly, it gets rid of the header that

contains space to fill in the name, unit, admission and discharge date. Then it ensures that the negations are treated properly and attached to the word they are negating in order to convey that meaning. Lastly, the special characters and extra spaces were removed but the numbers were kept.

Following the text cleaning, tokenization and lemmatization were performed. Tokenization involves the segmentation of clinical notes into individual tokens, either words or phrases, to facilitate further processing. This process breaks down the text into smaller units, making it easier to analyze. Lemmatization was then applied, reducing words to their base or root form. In order to obtain the numerical representations of the textual discharge while preserving medical contextual characteristics, we employed the Doc2Vec model [?]. This technique generates vectorized representations of documents by considering the relationships between words, capturing the semantic structures more effectively. The model was customized to generate 300 vectors. This vector size was chosen considering the mean and median length of our discharge notes (respectively 1,496 and 1,375 words). A vector size of 300 would capture enough semantic detail without excessive dimensionality, as well as allowing more room for rare words to be represented.

Once the embeddings were created, they were integrated into the dataset as structured numerical features. Now the dataset is enriched with numerical representations of the discharge notes that add meaning to the patients' stay and predictive modeling can be applied to test if the results improve with the textual information.

4.3 Model Hyper-Parameter Tuning and Calibration

Models were trained using 10-fold Cross-Validation (CV) on the training dataset to obtain the best parameters for each model [13]. The hyper-parameters for the five baseline models were also taken from the baseline study: For Logistic Regression, we tuned the inverse of the regularization strength (c) over the range $\{ 'C': 100, 10, 1.0, 0.1, 0.01 \}$. For Random Forest, we selected three hyper-

parameters to tune: the number of trees in the forest {10, 100, 1000}, the maximum number of features the algorithm considers at each split {'log2', 'sqrt'}, and the number of samples to split an internal node {2, 5, 10}. For Support Vector Machine we optimized the regularization parameter {1000, 100, 10, 1.0, 0.1} and the kernel coefficient, from {1, 0.1, 0.01, 0.001, 0.0001}. For XGBoost we tuned the boosting learning rate {0.1, 0.2, 0.3}, the number of boosting rounds {100, 200, 300, 400, 500}, and the maximum tree depth (between 2 to 10). Finally, for AdaBoost we varied the learning rate, the weight applied to each classifier at each boosting iteration, in {0.0001, 0.001, 0.01, 0.1, 1.0}, and the number of estimators at which the boosting is terminated in {10, 100, 1000}. Upon completion of training, we assessed model performance against the validation set using the following metrics: accuracy, precision, recall, F1-score, Brier-Score Loss, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Area Under the Precision-Recall Curve (AUC-PRC), Expected Calibration Error (ECE). After this evaluation, we calibrate the models using the Platt-Scaling (also known as sigmoid or logistic calibration) on 25% calibration data.

To assess the performance of the calibration, we only focused on two metrics [13]: Brier-Score Loss which measures both calibration and discrimination, and among those to measure calibration we choose ECE, as well as AUC-ROC to evaluate discrimination. We calculated ECE manually by binning the predictions from calibration_curve. The ECE score is determined by sorting the predictions and dividing them into k bins with an approximately equal number of patients per bin. It ranges from 0 to 1, with lower values indicating better calibration of the classifier.

5 Results

5.1 Data Exploration

In our initial dataset without discharge notes, we explored key patient information including demographics (age and sex), the length of stay from two ICU admissions, total and average doses of the top 68 medications, and indicators for 62 common comorbidities. One of the first things that stood out was the sparsity in many columns—meaning a lot of missing or zero values, especially in medication dosage features. When visualized, the plots clearly showed that most patients didn't receive certain medications like tramadol, trazodone, and benzodiazepines, while fluids like sodium chloride 0.9% flush and sterile water were used far more consistently. This pattern held for both sum doses and average doses, with some medications being widely administered and others used rarely, likely for specific or critical care needs. Understanding this sparsity is important—it influences how much a feature might contribute to a machine learning model. Highly sparse features might not add much predictive value and could be safely removed or grouped using techniques like Principal Component Analysis (PCA). Alternatively, domain expertise could help categorize medications into broader, more informative groups. Medications commonly used in daily care might remain, while rare-use drugs need to be weighed carefully before inclusion. Ultimately, sparsity analysis helps us decide which features to keep, combine, or transform to build a more efficient and interpretable predictive model.

5.2 Results from Paper

The Random Forest was the best-performing model in the study by Zhang and Kuo, yielding a test AUC of 0.716. The performance of XGBoost also gives a good record with the test AUC score of 0.712. The worst performance was achieved by this classifier: SVM, with an AUC of only 0.680,

practically the same value obtained from Logistic Regression, with just 0.001 of difference. The consistency and robustness of the model are further confirmed by the fact that the Random Forest model also performs the best out of the models during test and validation results, showing that the model is still able to predict well using unseen data. The study has also computed the lower and upper bound of 95% Confidence Intervals with the aim of determining the uncertainty of the average AUC. This metric shows that the lower 95% CI of RF is higher than the upper 95% of both Logistic Regression and SVM, confirming that in general Random Forest outperforms Logistic Regression and SVM. This calculation has also shown that AdaBoost and XGBoost performance is not far from the Random Forest since their 95% CI is higher than the lower bound of RF's 95% importance. The paper also details the best combination of hyperparameters for each classifier, in the case of Logistic Regression, the best inverse regularization strength was 100. In Random Forest, the best maximum number of features was set to the logarithm base 2 of the number of features in each tree, the better performing number of samples required to split an internal node resulted in 10, and lastly, the best number of trees in the forest was 1,000. In the case of SVM, the best inverse of regularization strength was 10, the best-performing kernel was the linear one with a coefficient of 1. For AdaBoost, it was found that the best-boosting iteration weight was 0.1, and the best maximum number of estimators was 1,000. And, for the last model, XGBoost performed at its best with a learning rate of 0.1, the best number of boosting rounds set to 400, and the best maximum depth of the tree is 3. It is also important to note that all five classifiers kept their discrimination capability after being calibrated, the ECE of all five classifiers is diminished after applying Platt scaling calibration, especially the AdaBoost classifier. Random Forest showed the best discrimination performance, with a high AUC before and after calibration was done. When the models were evaluated using a confusion matrix with a 0.5 decision threshold, Random Forest and AdaBoost had the highest sensitivity, which suggests they are much better at identifying true positive results. XGBoost had the highest specificity, which

means that it is more suitable for recognizing true negative outcomes. Random Forest has the highest F1-Score confirming that this model has the strongest overall performance of the binary classification.

5.3 Results from Paper extracted by us

From the study performed by Zhang and Kuo, Random resulted as the best-performing model, achieving a test AUC of 0.716. XGBoost also performed well, with the test AUC score of 0.712. The worst performance was achieved by SVM, with an AUC of only 0.680. The consistency and robustness of Random Forest were further confirmed by its consistency in performance across both test and validation results, showing that the model is able to generalize well to unseen data. To assess the model uncertainty, the study calculated 95% confidence intervals (CIs) for the average AUC scores. The lower 95% CI of RF is higher than the upper 95% of both Logistic Regression and SVM, confirming that in general, Random Forest outperforms Logistic Regression and SVM. This calculation has also shown that AdaBoost and XGBoost performance is not far from the Random Forest since their 95% CI is higher than the lower bound of RF's 95% importance. The study also determined the optimal hyperparameters for each classifier. For their best performing model, the best maximum number of features was set to the logarithm base 2 of the number of features in each tree, the better performing number of samples required to split an internal node resulted in 10, and lastly, the best number of trees in the forest was 1,000. All five classifiers kept their discrimination capability after calibration, with ECE scores improving after applying Platt scaling, especially the AdaBoost classifier. Random Forest showed the best discrimination performance, with a high AUC before and after calibration. The models were further evaluated using a confusion matrix with a 0.5 decision threshold: Random Forest and AdaBoost had the highest sensitivity, suggesting their strong performance in identifying true positive results. XGBoost exhibited the highest specificity, making it better suited for recognizing true negative outcomes. Random Forest also showed the highest F1-Score, confirming the strongest

overall performance of the binary classification across all classifiers.



Figure 2: AUC (ROC) Scores for Uncalibrated models

The best hyperparameters for each model were the following: for Logistic Regression, the best regularization strength was 100. For Random Forest, the best number of trees is 1000, the maximum features were set to the square root of the total features at each split, while the minimal sample size to split an internal node is 5. For XGBoost, the best parameters were a learning rate of 0.2, a maximum tree depth of 3, and 500 boosting rounds. For AdaBoost, the best learning rate was equal to 1.0, while the optimal number of estimators was 1000. For Naive Bayes, the optimal variance smoothing was set to $1e-5$. Compared to our reference study, the model which showed the best performance was XGBoost with an AUC score of 0.7313, while Random Forest closely followed the performance showing an AUC score of 0.7294. The worst-performing models was instead Naive Bayes, not exceeding 0.65 AUC scores.

Calibration Curve

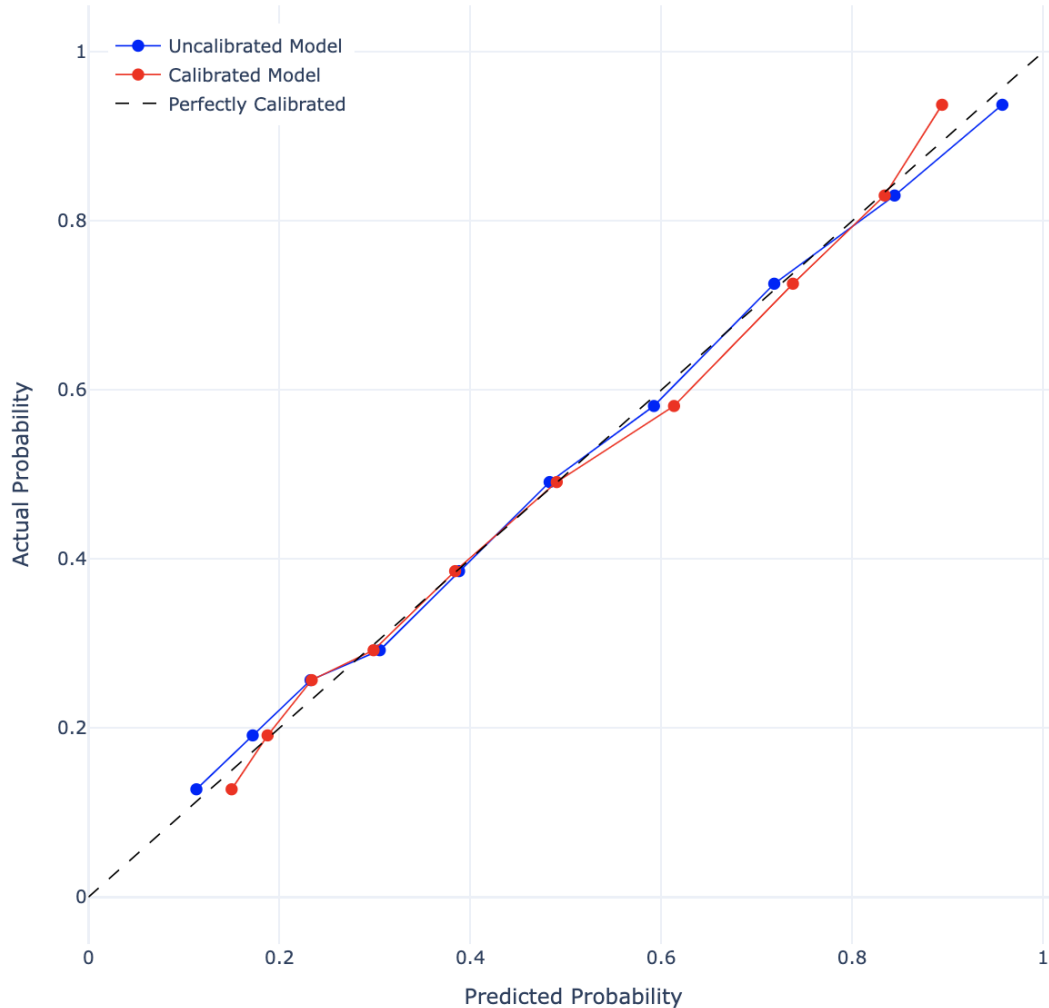


Figure 3: Calibration Curve

To better understand the performance of the models, confusion matrices were generated and analyzed. The results show that XGBoost has the highest accuracy among all models, meaning a better classification for most cases. This is further supported by its strong F1 score, demonstrating a good balance between false positives and false negatives. However, in terms of recall, Random Forest outperforms the other models, minimizing the misclassification of positive cases, which is particularly

important in medical cases.

Table 1: Calibration results for six classifiers. AUC: Area Under the Receiver Operating Characteristic (ROC) Curve; ECE: Expected Calibration Error.

Classifier	LR	RF	XGBoost	AdaBoost	SVC	NBC
Test AUC						
Before Calibration	0.6947	0.7294	0.7313	0.7054	0.7054	0.6418
After Calibration	0.6949	0.7295	0.7316	0.7119	0.6946	0.6436
Test ECE						
Before Calibration	0.0250	0.0407	0.0145	0.0679	0.0679	0.2253
After Calibration	0.0272	0.0184	0.0115	0.0203	0.0154	0.0928

5.4 Model results from an additional dataset with discharge notes

Across all models used, adding the discharge note embeddings has led to noticeable improvements in the performance metrics, notably accuracy, recall, and F1 score. The metrics which improved the most are recall and precision, indicating that the addition of the discharge notes helps the models to better capture true positive cases and maintain a balance in correct predictions, which is crucial in the medical field. It is important to note that the accuracy when modeling without discharge notes ranged from 0.65 to 0.73, while with discharge notes embeddings, the accuracy increased across all models, reaching 0.72 and 0.75. Recall also improved across the performance metrics, where XGBoost had the highest increase, from 0.6737 to 0.8248, an important result considering the context this study is performed, as misclassified instance could have a severe impact. F1-Score, the balance between precision and recall, increased significantly in XGBoost (from 0.7089 to 0.7804) making it the best-performing model out of all classifiers. AUC ROC and AUC PRC were also noticeably improved

when text embeddings were taken into account. The inclusion of the discharge notes as embeddings in the features of our dataset has proved to significantly improve predictive performance across all models, particularly in the recall, F1 Score, and AUC metrics. XGBoos consistently achieves the highest metrics, making it the most effective model even after adding the discharge note embeddings.

5.5 Feature Importance

After evaluating the results of every model used to classify whether the admission was going to be prolonged or not, it was important to understand which features were the ones contributing the most to the output in the models that were performing well. This allowed model interpretability and comprehension of what was driving the model's decision. For the Random Forest, the variables that most significantly impact the prediction of whether a second hospital admission will be prolonged are detailed in the graph below. The Mean Decrease in Impurity score represents how much each variable contributes to the improvement of the classification, the features that contain the higher value are the ones more relevant to the prediction. The total amount of Sodium Chloride 0.9% flush is the most important variable in the Random Forest, from this feature, we can infer that the patients who receive higher amounts of this medication may have specific medical conditions that require prolonged admissions. Digging up in the medical field and knowledge, it is found that this substance is used to clear intravenous lines, catheters, and ports to maintain their patency and prevent blockages, it ensures that medication administration remains smooth [12]. Understanding why this feature is relevant conveys a lot of important details of the medical field. Usually, patients who require this saline flush frequently are likely to be receiving continuous intravenous medication, hydration, or other critical treatments that may require extended hospital care, this flush also may suggest the presence of catheters which are commonly used for long-term administration of fluids, antibiotics or other treatments which are often correlated with longer stays [14]. The second significant predictor in

prolonged hospital admissions is the age of the patient, this is easy to understand since it is very likely that older patients experience complications that lead to longer hospital stays. This is also backed by research conducted in the medical field that shows that age is one of the strongest predictors of hospitalization duration [15] Another crucial factor to consider is how long the first hospital admission lasted, as this length may affect the second admission length. Patients who had a long period of admission on their first admission are most likely to have underlying health conditions which make for an increase in the risk of a prolonged second admission. Other important features are related to drug usage such as Bisacodyl, Ondansetron, Furosemide, etc. These medications are associated with gastrointestinal and cardiovascular diseases that may have a role in the prognostication of extended hospitalization.

- Bisacodyl signalizes gastrointestinal problems or pre-procedural preparation, which could indicate an underlying issue that requires extended medical care [16].
- Ondansetron is used to treat severe nausea, post-surgical recovery, or chemotherapy induced nausea, meaning its presence indicates patients who are going through intensive treatments [17].
- Furosemide relates to cardiovascular or kidney-related pathology, both well known risk factors for extended hospital stays [18].

After understanding what each medication is used for, it can easily be inferred that the presence of either of these, either as an average or sum, acts as an important factor when determining that admission is going to be prolonged.

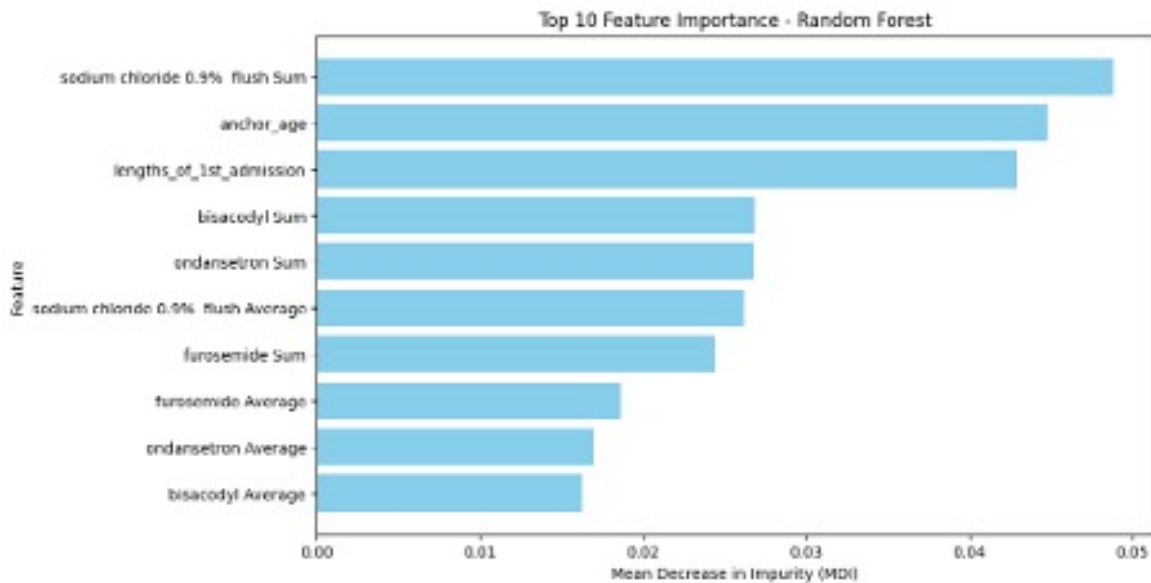


Figure 4: Top 10 Feature Importance- Random Forest

Analysis of feature importance was also conducted on the XGBoost model using Gain, which indicates how much each feature improves model splits. This model suggests that the second most important feature to predict a prolonged readmission is a recent childbirth, it could indicate postpartum complications that may require further hospital care. Medical studies on maternal health confirm that complications such as postpartum hemorrhage, infections and hypertension can lead to prolonged hospital stays [19]. The medication described in the feature importance of the Random Forest is also present in XGBoost and remains influential reinforcing their relevance. Acute kidney failure is also listed and easily aligns with a prolonged hospital stay, as kidney failure often leads to extended treatments, dialysis, and other crucial interventions [20]. A urinary tract infection is also relevant and can be understood since this type of condition can easily escalate to more severe infections like sepsis, this could also be relevant for older patients or those with chronic conditions [21].

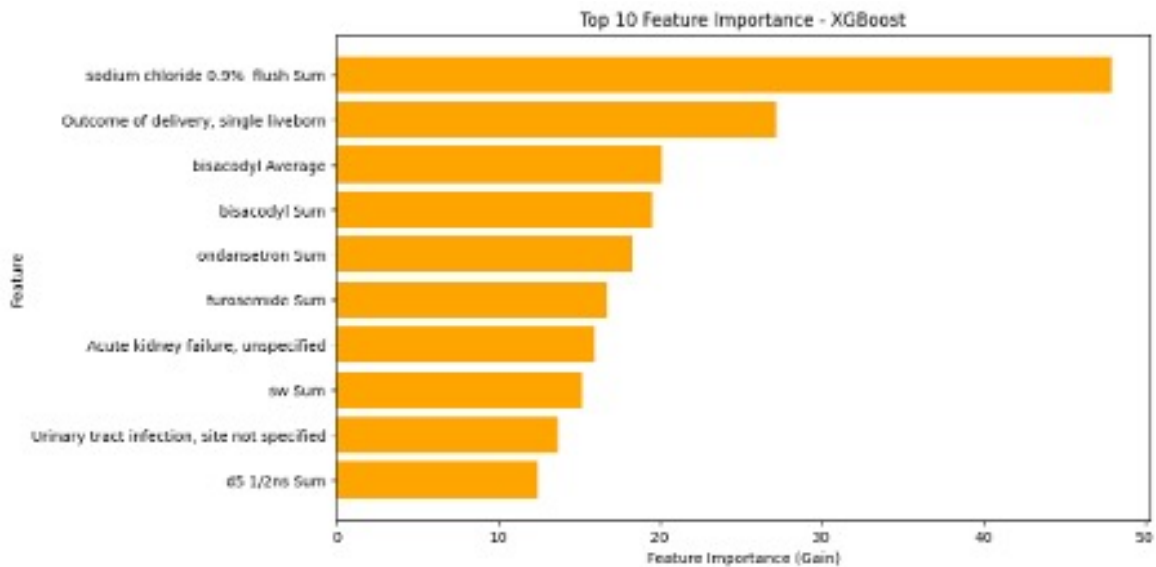


Figure 5: Top 10 Feature Importance- XGBoost Model

XGBoost places greater importance on diagnosis-related features like childbirth, kidney failure, and UTIs compared to the Random Forest model that emphasized age and first admission length but in both models the medication and treatment-related variables remain influential, reinforcing their importance. Each model picks up specific conditions that contribute to prolonged hospitalization, offering a different perspective on what increases a patient's readmission risk.

5.5.1 Feature Importance on discharge notes dataset

The feature importance returns small variations when it is performed on the dataset that includes the embeddings of the discharge notes. This information provides further interpretability and insights on additional features that influence the model's decision making. The Random Forest model highlights several factors aligned with the results obtained in the modeling of the previous dataset but tends to emphasize patient monitoring features like Emergency Unit Observation and Direct Observation. The XGBoost model, however, assigns higher importance to specific medical conditions and severity

scores. This includes the sum of the drug severity, a chronic obstructive pulmonary disease and a major depressive disorder, this provides a more diagnostic driven perspective that can be very useful. Both models include the sodium chloride, bisacodyl and furosemide, reinforcing their relevance in predicting prolonged admissions as previously stated by the embeddings less models.

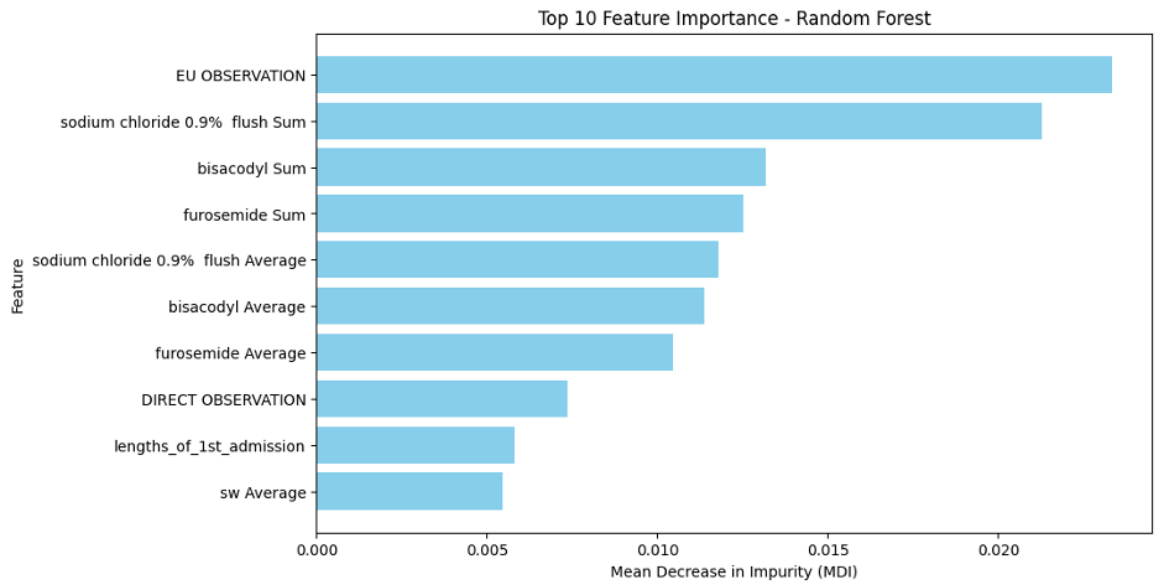


Figure 6: Top 10 Feature Importance- Random Forest- Dataset with Discharge Notes

6 Discussion

This study set out to reproduce and enhance the ICU length of stay (LOS) prediction framework proposed by Zhang and Kuo (2024) with the goal of identifying patients at risk of prolonged ICU admissions (≥ 3 days) for better hospital resource planning and management. Structured clinical data from the first ICU stay including medications, diagnoses, and discharge summaries, We were able to build machine-learning models. Among all the models, Random Forest and XGBoost stood out as top performers. These results are consistent with the findings of Zhang and Kuo (2024). Incorporating discharge note embeddings significantly improved recall and F1-score. This means discharge notes

have a good bearing on the second admission.

The most critical predictors are “Sodium Chloride 0.9% flush”, “length of the first ICU stay” and medications like Bisacodyl, Furosemide, and Ondansetron. XGBoost placed emphasis on recent childbirth, acute kidney failure, and urinary tract infection.

After model calibration using Platt scaling, performance improvement was observed across all classifiers through lower Brier scores and Expected calibration error.

Despite the promising results, our study faces narrow generalizability in terms that only patients with exact two admissions were chosen for modeling. Also, the incorporation of full ICU data will be more comprehensive and better at generalizing. Additionally, the noise inherent in discharge notes due to their unstructuredness may introduce inconsistencies in the predicted variable.

Nonetheless, the work highlights the credibility and worthiness of predictive analytics in critical care. Hospitals and Staff will always be short in the face of endemics or epidemics because of the sheer size of the population and share of caregiving professionals and infrastructure. The predictive tool will always be handy with respect to efficient resource utilization.

In the future, focusing on validating these models with external datasets pertaining to different hospitals to assess generalizability and incorporation of more dynamic and time-aware data and models would lead to betterment.

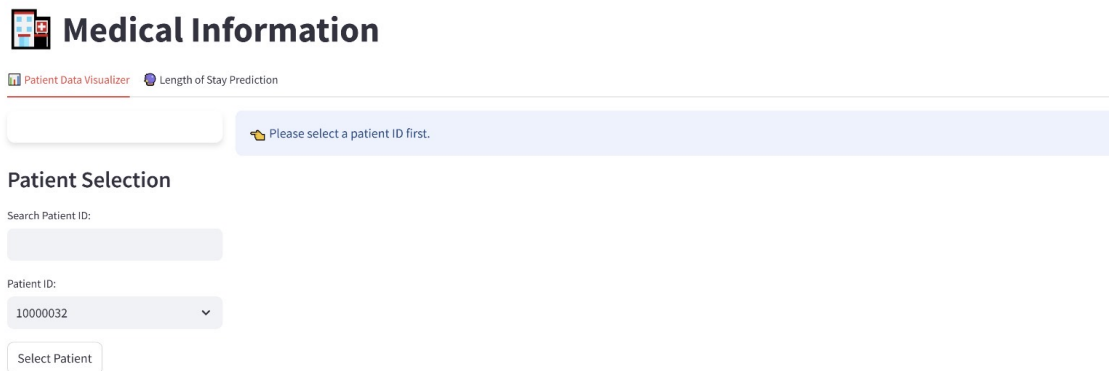
In essence, our study is at the juncture of critical care and machine learning. It places value on patient’s clinical history to predict future visits to the hospital. When combined with unstructured discharge notes, models can support smarter hospital operations.

7 Implementation (Streamlit)

To enable users to explore the data easily and to make our research more accessible at the same time, We created a lightweight web application using Streamlit, a Python-based open-source frame-


work for building interactive data apps. The main objective of the app was to provide a visual interface to explore and query the MIMIC-IV database, to better understand the database structure, semantics, and distribution of clinically relevant features.

Our Streamlit app connects to a local SQLite implementation of the MIMIC-IV database and enables the user to filter the data, through a responsive and easy-to-use interface, starting from selecting patient ID. After selecting a patient, all hospital admissions for a particular patient are retrieved, and at that time, the user should select a specific admission ID.





The screenshot shows the 'Medical Information' app interface. At the top, there's a header with a medical icon and the title 'Medical Information'. Below the header, there are two tabs: 'Patient Data Visualizer' (active) and 'Length of Stay Prediction'. A light blue banner with a warning icon and the text 'Please select a patient ID first.' is displayed. Under the 'Patient Selection' section, there's a 'Search Patient ID:' label above a text input field. Below that, a 'Patient ID:' label is above a dropdown menu showing '10000032'. At the bottom of this section is a 'Select Patient' button.

Then, once they click on it, the app presents a summary of that ICU stay. This will then display demographic information, including age, gender, and ethnicity, as well as clinically relevant information, such as diagnosis with ICD code labels, medication prescriptions, and dosages, laboratory test results, and all the information available from other relevant tables in the database, such as vital signs, procedures, and microbiology reports.



Medical Information

 Patient Data Visualizer

 Length of Stay Prediction

Search Patient ID:

10005348

Patient ID:

10005348

▼

Select Patient

Age

76 years

Race

WHITE

Language

ENGLISH

Gender

M

Marital Status

MARRIED

Insurance

Other

Admission Selection

Admission:

Admission Summary


Admission Type

EW EMER.

Admission Date

26/10/2130

An additional and significant feature of the application is the integration of our predictive model. For each patient admission displayed, the app also computes and presents a prediction of whether the selected ICU stay is likely to result in a prolonged length of stay (≥ 3 days). This prediction is based on the same machine learning models developed during our project and uses the corresponding features extracted from the patient's data as inputs. By embedding the model directly into the interface, we make it possible to simulate how such a prediction tool could be used in a real clinical environment to flag high-risk cases early in the admission process.



Medical Information

Patient Data Visualizer

Length of Stay Prediction

</

The interface is titled "Medical Information" with a medical icon. It features a navigation bar with "Patient Data Visualizer" and "Length of Stay Prediction". The main content is divided into four sections: "Patient Selection" with search and dropdown fields for Patient ID (10001338) and a "Select Patient" button; "Prediction Results" showing a green pill "Short ICU Stay (<3 days)"; "Clinical Insights" with a green message "Patient predicted to have a short ICU stay (<3 days) for second admission."; and "Admission Selection" with a "First Admission (used for prediction):" label and a list of "Clinical Considerations": "Standard care protocols likely sufficient" and "Consider early transition to step-down unit if appropriate".

This design enables users to dynamically, and more specifically, examine each hospital stay, allowing the reassembling of the greater clinical picture of their ICU stay. This application has meaningful implications not just for research and exploratory data analysis, but for the actual running of hospitals. If it provided an intuitive method for accessing and analyzing a patient’s own complete clinical history, then it could be a useful tool for ICU case review, early identification of high risk, and assistance with discharge planning or bed allocation. A hospital team, for instance, might apply it to track trends across admissions, analyze how medication pathways correlate with prolonged stays, or rapidly audit profiles of high-risk patients. The app highlights the potential to create interactive, visualization tools on the top clinical databases such as MIMIC-IV, which not only help in better decision-making but also lead to more efficient, data-driven healthcare management.

8 Conclusion

Using structured and unstructured data from MIMIC-IV v2, this study was conducted to predict whether a patient will have a prolonged second ICU stay (≥ 3 days). Using the framework from

Zhang and Kuo (2024), we added the demographic information, comorbidities, medication history, and the notes from discharge summaries to train interpretable machine learning models.

We found that risk factors including age, first admission length, and the use of specific medications significantly influenced LOS (length of stay). Random Forest and XGBoost performed consistently better than all other classifiers, especially in AUC and calibration metrics (Brier score). Inclusion of discharge note embeddings also improved predictive performance, most notably recall and F1-score. Visualizing the data in this way also led to the use of dimensionality reduction and clustering techniques which help interpret the features, since exploratory data analysis exhibited extreme sparsity in multiple features. Most importantly, our findings show that past information from the ICU is able to meaningfully predict future trajectories of the patient. Incorporating this model into a Streamlit app demonstrated how these types of predictive tools could be deployed in a clinical environment. From the view of business and hospital management, this effectively unlocks more intelligent ICU triage, smarter bed assignment, and more efficient care planning, showcasing the real-world possibilities of data-driven decision support in critical care.

9 Future Work

Analyzing the dataset, we see a lot of sparsity in the columns. Columns pertaining to the Sum and Average doses of top medications and Indicators of the presence of top 68 diagnoses are very sparse as shown in the EDA. One of the ways to handle sparsity is to use Principal Component Analysis. PCA would help reduce the dimension and simultaneously deal with the sparse nature of the dataset.

Also, the LOS threshold was selected in line with Zhang and Kuo (2024). Different thresholds can have a bearing on the model's performance. This can be tailored to specific hospitals or clinical settings.

Additionally, strategies for handling biases in terms of age, sex, and disease distribution need to be

addressed. Combining prediction of second admission with early hours information such as vitals information of second admission seems to be a more prudent and robust way of predicting the length of second admission.

In the near future, we might also see deep learning models trained on time series data of patient vitals and image signals.

List of Figures

1	Data Extraction Pipeline	6
2	AUC (ROC) Scores for Uncalibrated models	16
3	Calibration Curve	17
4	Top 10 Feature Importance- Random Forest	21
5	Top 10 Feature Importance- XGBoost Model	22
6	Top 10 Feature Importance- Random Forest- Dataset with Discharge Notes	23
7	AUC (ROC) Scores for Uncalibrated models	VI
8	AUC (ROC) Scores for Uncalibrated models	VII
9	AUC (ROC) Scores for Uncalibrated models	VIII
10	AUC (ROC) Scores for Uncalibrated models	IX
11	AUC (ROC) Scores for Uncalibrated models	X
12	AUC (ROC) Scores for Uncalibrated models	XI
13	AUC (ROC) Scores for Uncalibrated models	XII
14	AUC (ROC) Scores for Uncalibrated models	XIII
15	AUC (ROC) Scores for Uncalibrated models	XIV
16	AUC (ROC) Scores for Uncalibrated models	XV
17	AUC (ROC) Scores for Uncalibrated models	XVI
18	AUC (ROC) Scores for Uncalibrated models	XVII

List of Tables

1 Calibration results for six classifiers. AUC: Area Under the Receiver Operating Characteristic (ROC) Curve; ECE: Expected Calibration Error. 18



References

- [1] M. Zhang and T.-T. Kuo, “Early prediction of long hospital stay for intensive care units readmission patients using medication information,” *Computers in Biology and Medicine*, vol. 174, p. 108451, 2024.
- [2] M. Syed, S. Syed, K. Sexton, H. B. Syeda, M. Garza, M. Zozus, F. Syed, S. Begum, A. U. Syed, J. Sanford, and F. Prior, “Application of machine learning in intensive care unit (icu) settings using mimic dataset: Systematic review,” *Informatics*, vol. 8, no. 1, p. 16, 2021.
- [3] M. Gupta, N. Gallamoza, N. Cutrona, P. Dhakal, R. Poulain, and R. Beheshti, “An extensive data processing pipeline for mimic-iv,” *arXiv preprint*, 2022.
- [4] S. R. Khope and S. Elias, “Critical correlation of predictors for an efficient risk prediction framework of icu patient using correlation and transformation of mimic-iii dataset,” *Data Science and Engineering*, vol. 7, no. 1, pp. 71–86, 2022.
- [5] X. Wang, Y. Li, and Q. Zhang, “Establishment of icu mortality risk prediction models with machine learning algorithm using mimic-iv database,” *Diagnostics*, vol. 12, no. 5, p. 1068, 2022.
- [6] T. Lalanne and S. Dupont, “Evaluating the fairness of the mimic-iv dataset and a baseline algorithm: Application to the icu length of stay prediction,” *arXiv preprint*, 2024.
- [7] J. B. Benzine and J. Lee, “Mimic-iv-icd: A benchmark for extreme multi-label classification,” *arXiv preprint*, 2023.
- [8] F. Jaotombo, L. Adorni, B. Ghattas, and L. Boyer, “Finding the best trade-off between performance and interpretability in predicting hospital length of stay using structured and unstructured data,” *PLOS ONE*, vol. 18, pp. 1–22, 11 2023.

-
- [9] M. Edelson and T.-T. Kuo, “Generalizable prediction of covid-19 mortality on worldwide patient data,” *JAMIA Open*, vol. 5, p. ooac036, 05 2022.
- [10] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, “A simple algorithm for identifying negated findings and diseases in discharge summaries,” *Journal of Biomedical Informatics*, vol. 34, no. 5, pp. 301–310, 2001.
- [11] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” 2014.
- [12] S. Sánchez-Román and T. L. Smith, “The role of sodium chloride flushes in catheter maintenance,” *Journal of Critical Care Nursing*, vol. 45, no. 3, pp. 215–230, 2021.
- [13] Y. Huang, W. Li, F. Macheret, R. A. Gabriel, and L. Ohno-Machado, “A tutorial on calibration measurements and calibration models for clinical prediction models,” *Journal of the American Medical Informatics Association*, vol. 27, pp. 621–633, 02 2020.
- [14] N. P. O’Grady and M. Alexander, “Guidelines for the prevention of intravascular catheter-related infections,” *Clinical Infectious Diseases*, vol. 52, no. 9, pp. e162–e193, 2011.
- [15] M. L. Barnett and J. Hsu, “The influence of age on hospital readmissions,” *New England Journal of Medicine*, vol. 382, no. 8, pp. 764–773, 2020.
- [16] X. Yao and L. Ge, “Gastrointestinal disorders and medication use,” *Gastroenterology Journal*, vol. 45, no. 3, pp. 215–230, 2019.
- [17] J. Smith and N. Patel, “The role of ondansetron in hospitalized patients,” *Clinical Pharmacology and Therapeutics*, vol. 104, no. 3, pp. 527–536, 2018.
- [18] F. H. Messerli and S. Bangalore, “Diuretics and heart failure,” *European Heart Journal*, vol. 38, no. 7, pp. 515–523, 2017.
-

-
- [19] W. M. Callaghan and A. A. Creanga, “Maternal mortality and severe morbidity in the united states,” *Obstetrics and Gynecology*, vol. 132, no. 1, pp. 60–69, 2018.
- [20] C. Y. Hsu and G. M. Chertow, “Chronic kidney disease and the risks of hospitalization,” *Kidney International*, vol. 95, no. 4, pp. 727–739, 2019.
- [21] B. Foxman, “Urinary tract infection syndromes,” *Clinical Infectious Diseases*, vol. 69, no. 2, pp. 211–217, 2019.

Appendix: Exploratory Data Analysis

In the initial dataset, we have columns pertaining to demographics – Age, Sex, Two lengths of stays – First and second admission, Sum and Average doses of the top 68 medications, and Binary indicators of the top 62 comorbidities. We went through them analyzing them one by one. The first notable characteristic of the dataset is sparsity in many columns.

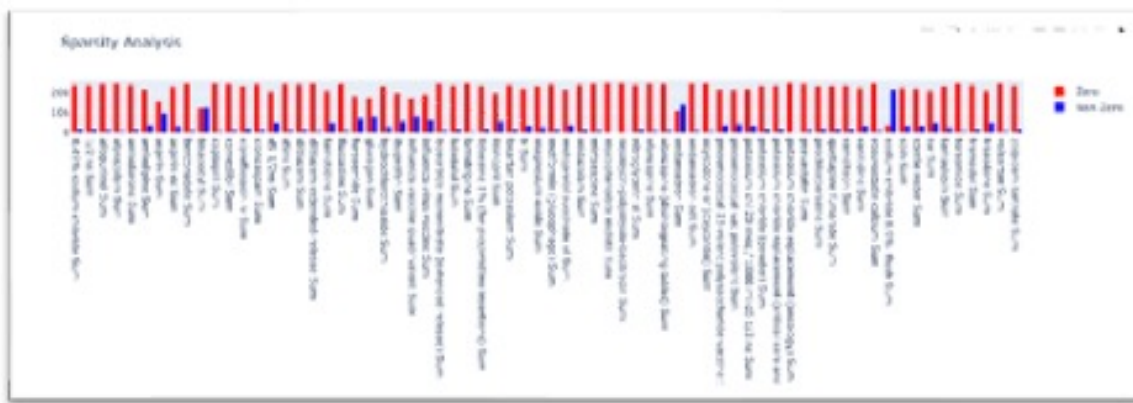


Figure 7: AUC (ROC) Scores for Uncalibrated models

The plot provides a peek into the sparsity of the columns corresponding to summed sum dose of various medications received by the patients. The red shows the zero counts while blue shows the non-zero values. X-axis is for different columns while Y-axis represents the count. The dominance of red color clearly states that columns are very sparse. Tall red bars corresponding to "tramadol Sum", "tramadol Sum", and "benzodiazepine Sum", tells us that most patients did not receive these medications. Some dominant blue bars can also be seen, for example, "sodium chloride 0.9% flush Sum" and "sterile water Sum" indicating these are common use fluids/flushing agents.

Sparsity analysis is important because it might play a role in feature selection for machine learning models. Because contribution to the ML model is inversely proportional to sparsity. However, correlation with the target variable might still give value to them. Another thing is the application of

techniques like PCA would be able to reduce the dimensions of such columns if the model needs to be deployed and scaled. Sparsity also has an implication on the rarity of the medication administered only to patients with special needs. Low-sparsity medications are more in routine care. One possible way out is to use domain knowledge and combine medications into broader categories and sum them to reduce dimensionality. Certain medications might be critical and weighting them to improve their contribution to the model might be needed but that requires domain expertise.

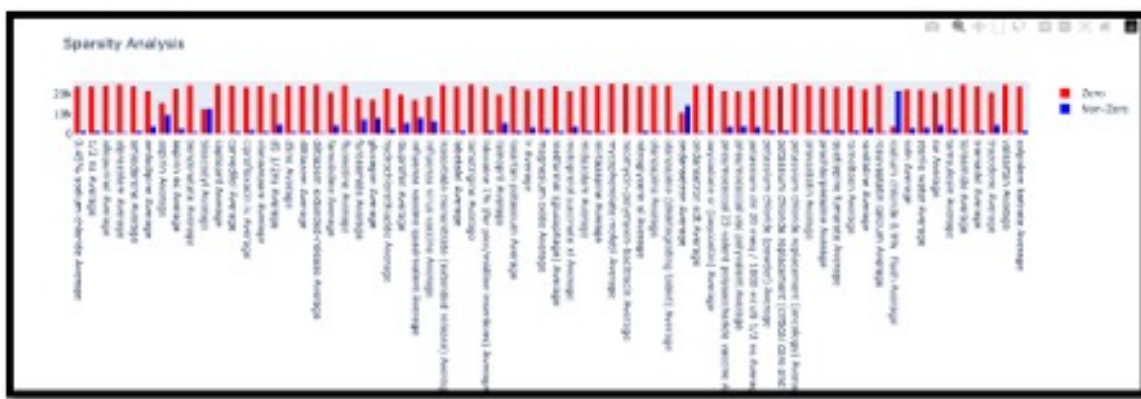


Figure 8: AUC (ROC) Scores for Uncalibrated models

Similar to Sum dose columns, Average dose columns also present sparse behavior. The dominance of red color is witness to this. Columns like “trazodone Average”, and “tramadol Average” have very tall red bars indicating they are mostly sparse having been administered to only a few patients. On the other hand, Columns such as “sodium chloride 0.9% flush Average”, which are used on a daily basis for general care, are administered to every patient and mostly show very little sparsity. If the correlation with length of 2nd admission is less, The sparsity in these columns will make their contribution negligible.

High sparsity in both the "Sum" and "Average" columns reinforces the idea that certain medications may not contribute significantly to predictive models. Features with extremely high sparsity could be considered for PCA to reduce dimensionality, especially if they do not show strong correla-

tions with outcomes.

.1 Univariate Analysis

.1.1 Age Distribution

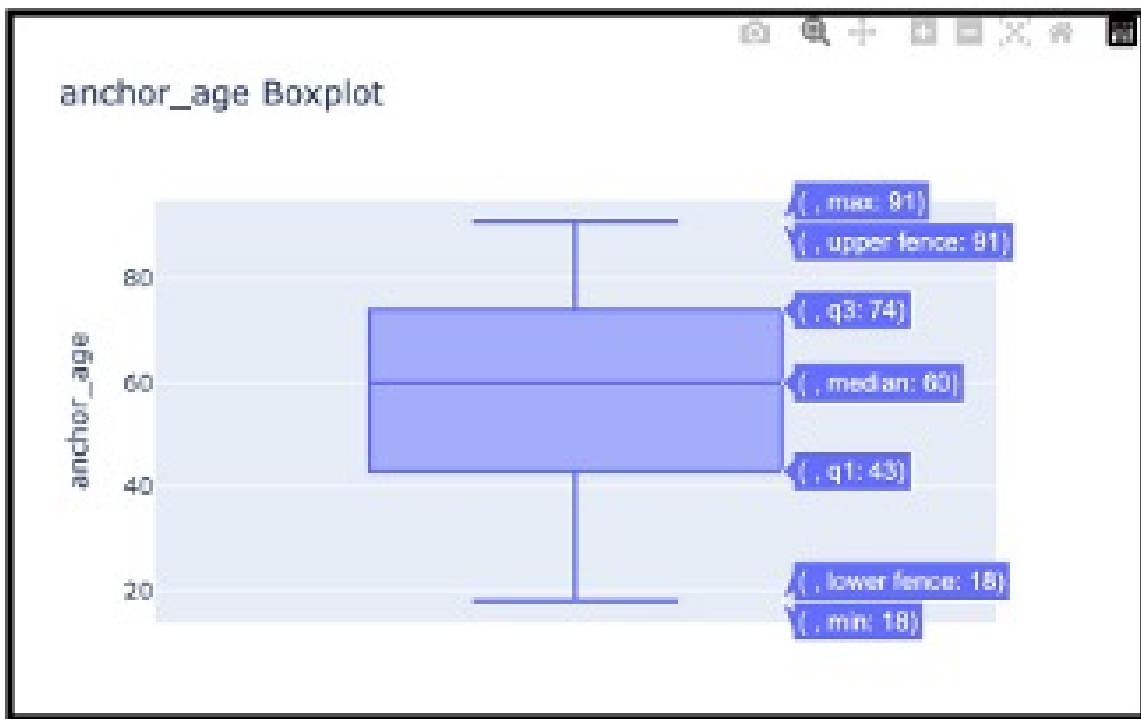


Figure 9: AUC (ROC) Scores for Uncalibrated models

The first parameter in the analysis is Age. Age is a critical demographic variable that can influence length of stay outcomes. The minimum age in the dataset is 18 years. The first quartile (Q1) is at 43 years, meaning that 25% of the patients are 43 years old or younger. The median age is 60 years indicating that half of the patients are 60 years old or younger and the other half are 60 years old or older. The third quartile (Q3) is at 74 years, indicating that 75% of the patients are 74 years old or younger. The maximum age in the dataset is 91 years. There are more younger patients compared to

very elderly patients.

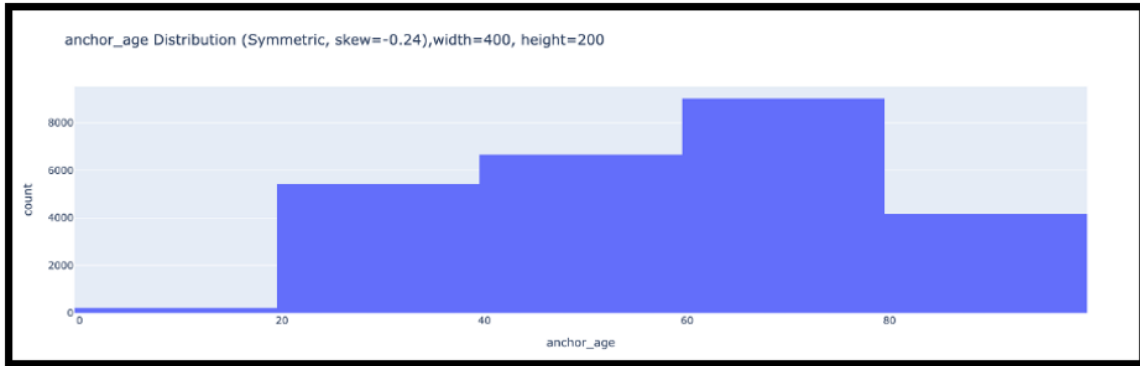


Figure 10: AUC (ROC) Scores for Uncalibrated models

The histogram appears to be approximately symmetric with a slight left skew. The peak of the distribution is centered around the early 60s. The majority of the data points are concentrated between 30 and 80 years, with fewer observations at the extremes. This aligns with the median age observed in the boxplot (60 years), further confirming that the central tendency of the age distribution is around this range. Given the concentration of patients, it may be important to focus on this age group when analyzing trends or making predictions.

.1.2 Length of 1st Admission distribution

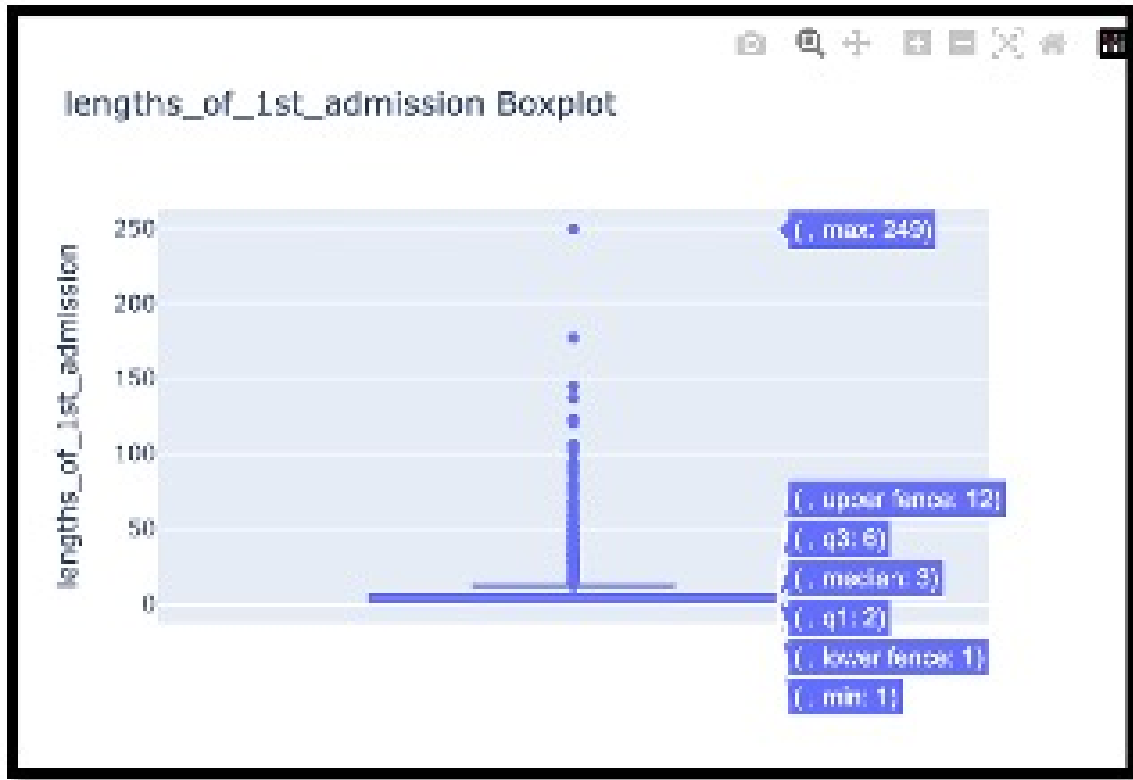


Figure 11: AUC (ROC) Scores for Uncalibrated models

The length of 1st admission boxplot shows a highly right-skewed distribution. The minimum length of the first admission is 1 day. The first quartile (Q1) is at 2 days. The median length of the first admission is 3 days indicating that half of the patients had a first admission lasting 3 days or fewer. The third quartile (Q3) is at 6 days. The maximum length of the first admission is 249 days which is an extreme outlier.

.1.3 Length of 2nd Admission distribution



Figure 12: AUC (ROC) Scores for Uncalibrated models

For the length of 2nd admission, A value of 0 indicates that some patients did not have a second admission or had a very short stay. The first quartile (Q1) is at 1 day. The median length of the second admission is 2 days. The third quartile (Q3) is at 5 days meaning that 75% of the patients had a second admission lasting 5 days or fewer. The maximum length of the second admission is 200 days, which is a significant outlier.

The boxplot shows a highly right-skewed distribution. The median (2 days) is closer to the lower end of the interquartile range (IQR), and the upper whisker extends much further than the lower whisker. This indicates that most patients had relatively short second admissions.

.1.4 Disease Prevalence

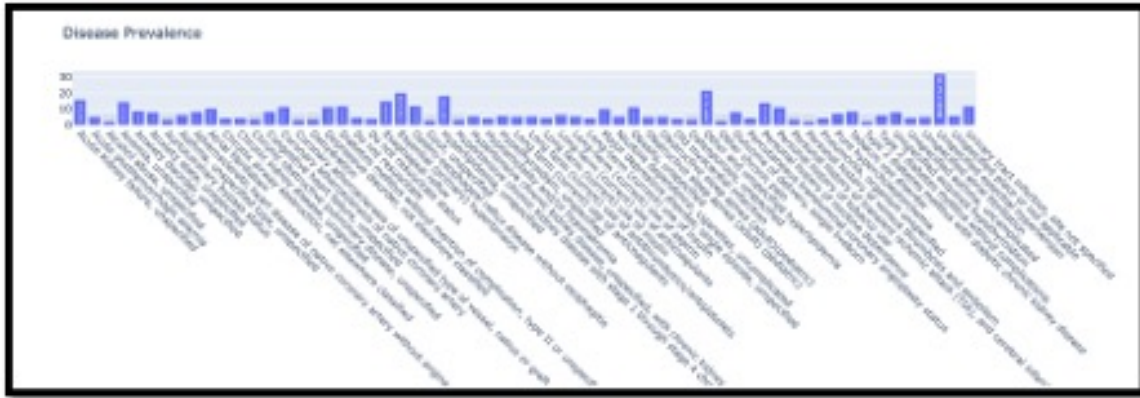


Figure 13: AUC (ROC) Scores for Uncalibrated models

We also tried to find out which disease is most prevalent in the dataset. The x-axis lists different diseases. The y-axis represents the count or frequency of patients diagnosed with each disease. Diseases like “Atrial fibrillation”, “Chronic obstructive pulmonary disease” and “Hypertension” appear to have high counts. Many diseases have a very low prevalence, as indicated by shorter bars. Acute conditions such as sepsis, respiratory failure, and acute kidney injury are also prominent indicating that the dataset includes many critically ill patients. High-prevalence diseases (e.g., hypertension, diabetes) are likely to be strong predictors of patient outcomes.

.2 Bivariate Analysis

.2.1 Age V/s Gender

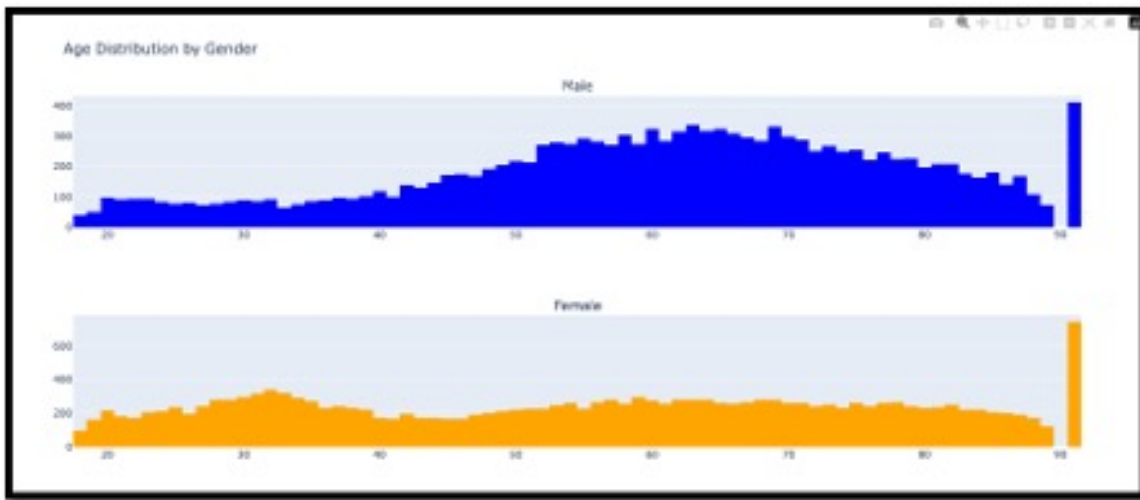


Figure 14: AUC (ROC) Scores for Uncalibrated models

The x-axis represents age (in years), and the y-axis represents the count of patients within each age bin. The distribution for males is right-skewed, with a peak around the early 60s. There is a gradual increase in the number of patients from younger ages (20–40 years) to the peak, followed by a slow decline as age increases.

The distribution for females is also right-skewed but appears slightly more spread out compared to males. Similar to males, there is a gradual increase in the number of patients from younger ages (20–40 years) to a peak, followed by a slow decline. The peak for females is occurring in the early 50s. Females show a broader distribution compared to males, with a more even spread across the age bins.

.2.2 Age V/s Disease Correlation

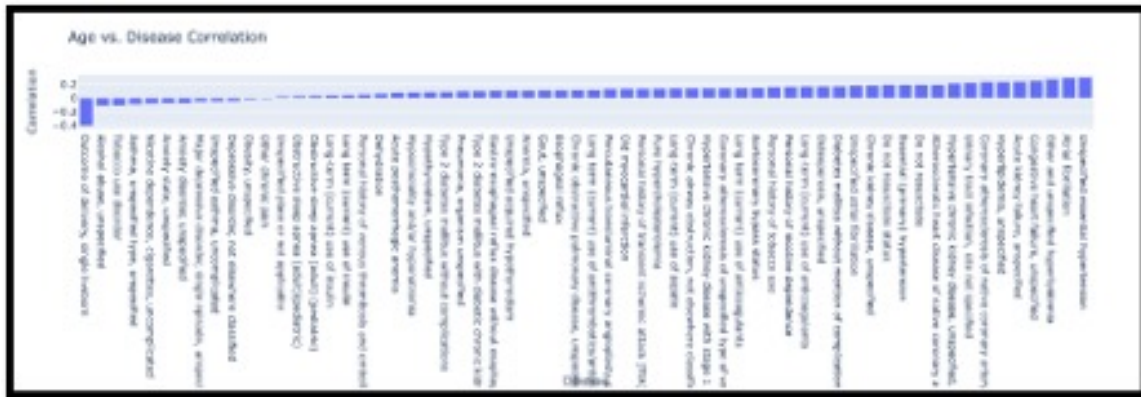


Figure 15: AUC (ROC) Scores for Uncalibrated models

The x-axis lists different diseases. The y-axis represents the correlation coefficient, ranging from approximately -0.2 to 0.2. Each bar indicates the strength and direction of the correlation between age and the corresponding disease. A positive correlation suggests that the prevalence of the disease increases with age. A negative correlation suggests that the prevalence of the disease decreases with age.

Several diseases show positive correlations with age, indicating that their prevalence increases as patients get older. A strong positive correlation suggests that COPD becomes more prevalent with advancing age. Another significant positive correlation, reflecting the well-known association between hypertension and aging. A moderate positive correlation, consistent with the known increase in diabetes risk with age. A positive correlation, aligns with the fact that cardiovascular diseases are more common in older populations. A positive correlation, as heart failure, is often a consequence of chronic conditions that develop over time. Few diseases show negative correlations with age, indicating that their prevalence decreases as patients get older. A slight negative correlation suggests that acute kidney injury is less common in older patients compared to younger ones. A weak negative

correlation, possibly due to the higher incidence of gastrointestinal issues in younger populations. A near-zero correlation implies that anxiety disorders are not significantly influenced by age. A near-zero correlation indicates that asthma prevalence is relatively stable across age groups.

.2.3 Disease V/s Length of 1st admission

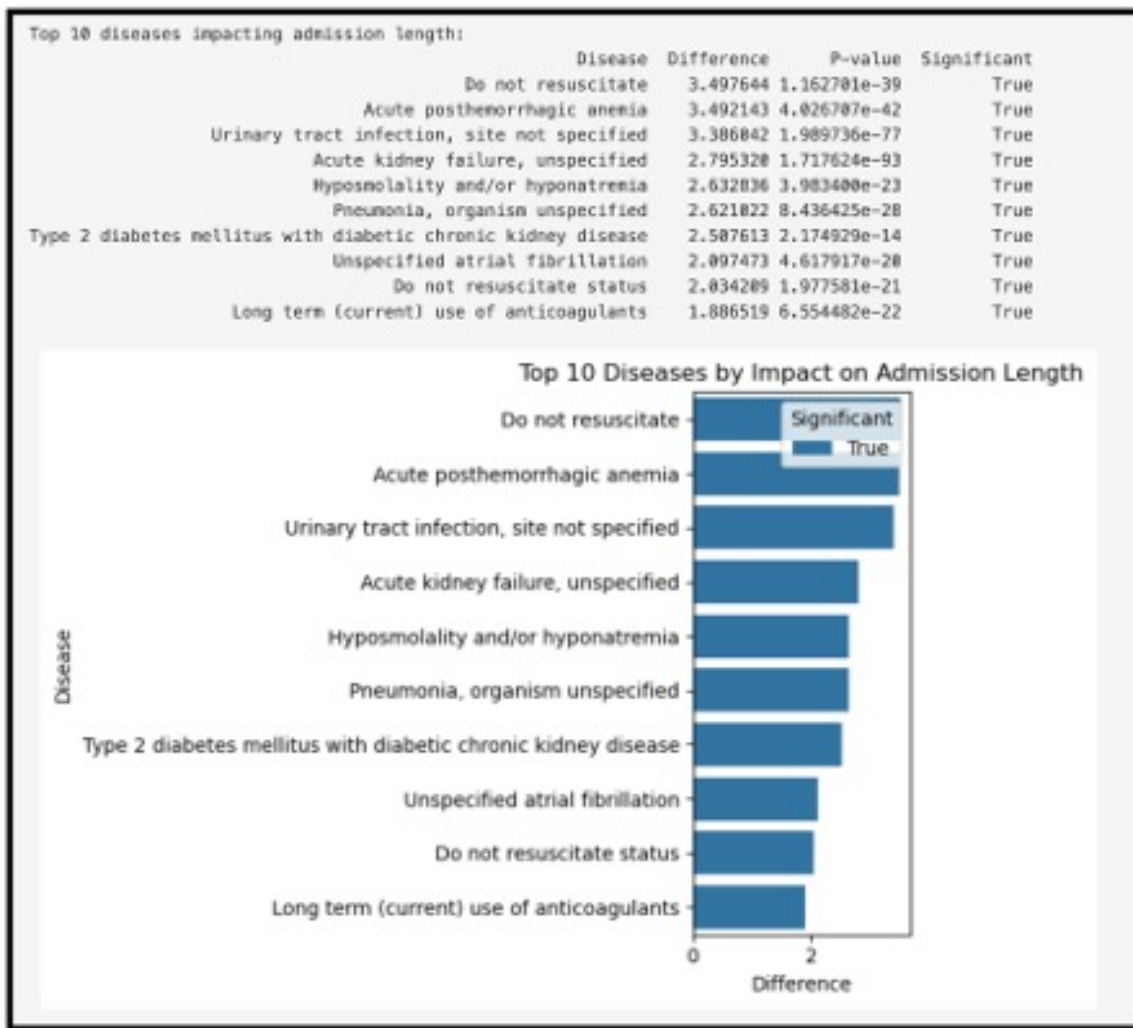


Figure 16: AUC (ROC) Scores for Uncalibrated models

The bar chart visually represents the top 10 diseases by their impact on admission length. The x-axis represents the difference in admission length (in days). The y-axis lists the diseases. Each bar corresponds to a disease, with the length of the bar indicating the magnitude of the difference.

"Do not resuscitate" has the highest impact, with a difference of approximately 3.5 days. This is followed closely by "Acute posthemorrhagic anemia" and "Urinary tract infection, site not specified," both with differences around 3.4 days. Diseases like "Do not resuscitate," "Acute posthemorrhagic anemia," and "Urinary tract infection, site not specified" have the most substantial impact on admission length. These conditions likely require prolonged hospital stays due to their severity, complexity, or need for intensive care. The analysis reveals that certain diseases, such as "Do not resuscitate," "Acute posthemorrhagic anemia," and "Urinary tract infection, site not specified," have a significant and substantial impact on hospital admission length.

.2.4 Top 10 Correlations between Age, Length V/s Medications

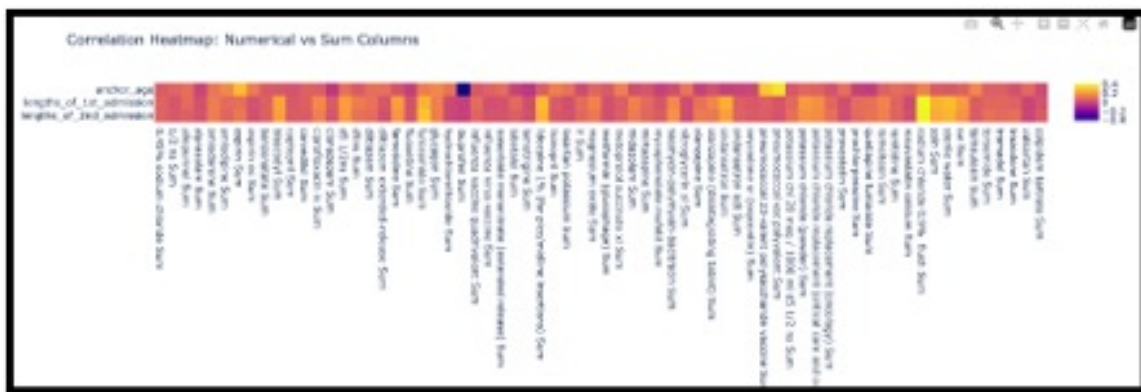


Figure 17: AUC (ROC) Scores for Uncalibrated models

The negative correlation of Age Vs Ibuprofen- It suggests that as a patient's age increases, the total dose of ibuprofen they receive tends to decrease. "lengthsof1stadmission" vs "sodium chloride 0.9% flush Sum". A strong positive correlation indicates that longer first admissions are associated

with higher total doses of sodium chloride 0.9% flush. Sodium chloride 0.9% flush is commonly used which may be required more frequently during longer hospital stays. Similar to the first correlation, this suggests that longer second admissions are also associated with higher total doses of sodium chloride 0.9% flush. “anchorage” vs “pneumococcal vac polyvalent Sum”: 0.366. As age increases, the total dose of pneumococcal vaccine received also increases. Older patients are more likely to receive pneumococcal vaccines due to their higher risk of pneumonia and other infections. “anchorage” vs “pneumococcal 23-valent polysaccharide vaccine Sum”: 0.310. Similar to the previous correlation, this reflects an association between age and vaccination. “lengthsof2ndadmission” vs “furosemide Sum”: 0.301. Longer second admissions are associated with higher total doses of furosemide. “anchorage” vs “aspirin Sum”: 0.289. As age increases, the total dose of aspirin received also increases. Older patients are more likely to use aspirin for cardiovascular disease prevention or management. “lengthsof1stadmission” vs “lidocaine 1% (for PICC/midline insertions) Sum”. Longer first admissions are associated with higher total doses of lidocaine. “lengthsof1stadmission” vs “sterile water Sum”: 0.284. Longer first admissions are associated with higher total doses of sterile water. ‘lengthsof2ndadmission’ vs ‘sterile water Sum’: 0.268. Similar to the previous correlation, this suggests that longer second admissions are associated with higher total doses of sterile water.

.2.5 Top 10 Correlated Disease Pairs

Top correlated disease pairs:
 Chronic kidney disease, unspecified <=> Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage 1 through stage IV, or unspecified: Correlation = 0.76
 Arteriosclerotic disease status <=> Coronary atherosclerosis of unspecified type of vessel, native or graft: Correlation = 0.75
 Coronary atherosclerosis of native coronary artery <=> Percutaneous transluminal coronary angioplasty status: Correlation = 0.55
 Essential (primary) hypertension <=> Hyperlipidemia, unspecified : Correlation = 0.49
 Atrial fibrillation <=> Long-term (current) use of anticoagulants: Correlation = 0.48
 Anxiety disorder, unspecified <=> Major depressive disorder, single episode, unspecified: Correlation = 0.45
 Coronary atherosclerosis of native coronary artery <=> Old myocardial infarction : Correlation = 0.44
 Other and unspecified hyperlipidemia <=> Unspecified essential hypertension: Correlation = 0.42
 Old myocardial infarction <=> Percutaneous transluminal coronary angioplasty status: Correlation = 0.42
 Hypertensive chronic kidney disease with stage 1 through stage 4 chronic kidney disease, or unspecified chronic kidney disease <=> Type 2 diabetes mellitus with diabetic chronic kidney disease

Figure 18: AUC (ROC) Scores for Uncalibrated models

Chronic kidney disease, unspecified – Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage I through stage IV, or unspecified: Correlation = 0.76. This is the strongest correlation in the list.

Aortocoronary bypass status – Coronary atherosclerosis of unspecified type of vessel, native or graft: Correlation = 0.75. A strong correlation indicates that patients with coronary artery bypass grafts (CABG) are highly likely to have coronary atherosclerosis.

Coronary atherosclerosis of native coronary artery – Percutaneous transluminal coronary angioplasty (PTCA) status: Correlation = 0.51

Coronary atherosclerosis of native coronary artery – Old myocardial infarction: Correlation = 0.44. Atherosclerosis is a major cause of heart attacks, making this correlation clinically intuitive.

Old myocardial infarction – Percutaneous transluminal coronary angioplasty status: Correlation = 0.41. Patients with a history of myocardial infarction are more likely to undergo angioplasty to prevent further cardiac events.

Hypertension and Hyperlipidemia - Essential (primary) hypertension – Hyperlipidemia, unspecified: Correlation = 0.49. Hypertension and hyperlipidemia are strongly correlated, reflecting their shared role as risk factors for cardiovascular diseases.

Other and unspecified hyperlipidemia – Unspecified essential hypertension: Correlation = 0.42. Similar to the previous correlation, this reflects the association between lipid disorders and high blood pressure.

Atrial fibrillation – Long-term (current) use of anticoagulants: Correlation = 0.49. Atrial fibrillation (AF) is strongly associated with anticoagulant use, as these medications are prescribed to reduce the risk of stroke in AF patients. Atrial fibrillation – Congestive heart failure, unspecified: Correlation = 0.44. AF and congestive heart failure (CHF) often coexist due to shared risk factors (e.g., aging, hypertension) and mutual exacerbation. Anxiety disorder, unspecified – Major depressive

disorder, single episode, unspecified: Correlation = 0.43. Anxiety and depression frequently co-occur, reflecting their overlapping symptomatology and shared neurobiological pathways.