Assessment and Prediction of Clinical Outcomes for ICU-Admitted Patients Diagnosed with Hepatitis: Integrating Sociodemographic and Comorbidity Data.

Dimple Sushma Alluri[1], and Felix M. Pabon-Rodriguez[1,2]*

[1]Department of Biomedical Engineering and Informatics, Indiana University, Luddy School of Informatics, Computing, and Engineering, Indianapolis, Indiana, United States

[2]Department of Biostatistics and Health Data Science, Indiana University, School of Medicine, Indianapolis, Indiana, United States

Correspondence
*Felix M. Pabon-Rodriguez
Email: fpabonrodriguez@gmail.com

Abstract

Hepatitis, a leading global health challenge, contributes to over 1.3 million deaths annually, with hepatitis B and C accounting for the majority of these fatalities. Intensive care unit (ICU) management of patients is particularly challenging due to the complex clinical care and resource demands. This study focuses on predicting Length of Stay (LoS) and discharge outcomes for ICU-admitted hepatitis patients using machine learning models. Despite advancements in ICU predictive analytics, limited research has specifically addressed hepatitis patients, creating a gap in optimizing care for this population. Leveraging data from the MIMIC-IV database, which includes around 94,500 ICU patient records, this study uses sociodemographic details, clinical characteristics, and resource utilization metrics to develop predictive models. Using Random Forest, Logistic Regression, Gradient Boosting Machines, and Generalized Additive Model with Negative Binomial Regression, these models identified medications, procedures, comorbidities, age, and race as key predictors. Total LoS emerged as a pivotal factor in predicting discharge outcomes and location. These findings provide actionable insights to improve resource allocation, enhance clinical decision-making, and inform future ICU management strategies for hepatitis patients.

KEYWORDS: Hepatitis, predictive models, resource utilization, health disparity, SDOH.

1. INTRODUCTION

Hepatitis, an inflammatory liver disease, remains a significant global health challenge, claiming an increasing number of lives each year. According to the World Health Organization (WHO) 2024 Global Hepatitis Report, viral hepatitis is the second leading infectious cause of death globally, responsible for 1.3 million deaths annually. This number has risen from 1.1 million in 2019, with 83% of these deaths attributed to hepatitis B and 17% to hepatitis C (WHO, 2024). Every day, approximately 3,500 people die due to hepatitis B and C infections worldwide. Despite advancements in diagnostic tools and treatment options, testing and treatment coverage rates have plateaued, signaling a growing public health crisis (WHO, 2024). In the United States, the most common forms of viral hepatitis are hepatitis A, B, and C, each impacting the liver differently and predominantly affecting distinct populations (CDC, 2024). Hepatitis B and C pose severe health risks, often leading to chronic conditions such as cirrhosis and liver cancer. These diseases are also the primary contributors to liver-related mortality globally (WHO, 2023). Managing intensive care unit (ICU) - admitted patients with hepatitis is particularly challenging due to the treatment complexities and the resource-intensive nature of care required for severe cases.

One critical issue in ICU settings is the unpredictability of a patient's length of stay (LoS), a key metric influencing hospital resource management, patient care quality, and healthcare efficiency. Prolonged ICU stays are closely linked to increased hospital costs and heightened resource strain (Peres et al., 2020).

Research also indicates that longer ICU stays correlate with increased long-term mortality rates, underscoring the need for precise LoS predictions (Moitra et al., 2015). Beyond LoS, discharge outcomes and locations serve as vital metrics for assessing patient recovery, readmission risks, and the burden on healthcare systems (Hickman, 2018). Sociodemographic and clinical factors—including race, gender, marital status, insurance type, age, and type of hepatitis—significantly influence hospitalization outcomes, particularly for ICU-admitted hepatitis patients. These factors also play a crucial role in determining hospital stay duration and discharge outcomes (Arnab Kumar Ghosh et al., 2021; Hayes et al., 2016; Dubin et al., 2024; Eskandari et al., 2022; Ng et al., 2024).

Recent years have seen the growing application of machine learning (ML) in healthcare, particularly for predictive modeling. ML algorithms provide several advantages over traditional statistical approaches, including their ability to analyze large, complex datasets and identify subtle patterns that conventional techniques may overlook (An et al., 2023). Predictive models using ML have been successfully applied in ICU settings to improve forecasts related to patient outcomes, treatment responses, and resource needs (Choi et al., 2022; Levin et al., 2020). However, there is a noticeable lack of ML models tailored to hepatitis patients and their unique clinical profiles.

The objective of this research is to explore and develop ML models to predict the length of stay, discharge location, and discharge outcomes for ICU-admitted hepatitis patients. These models will incorporate sociodemographic and clinical variables from ICU admission records as key predictors. By addressing this gap, the study aims to provide actionable insights for healthcare providers, enhancing resource planning and improving hepatitis patient outcomes in ICU settings.

2. MATERIALS AND METHODS

2.1. Data

The data set used for this study was sourced from the Medical Information Mart for Intensive Care (MIMIC-IV), version 3.0, a comprehensive and deidentified repository of patient records (Johnson et al., 2024; Johnson et al., 2023; Goldberger et al., 2000). MIMIC-IV comprises data collected from patients admitted to the emergency department (ED) or intensive care units at the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA. It includes over 364,000 unique patient records, with a total of 546,000 hospital admissions and nearly 94,500 ICU stays, providing a rich source of information for developing predictive models in healthcare. MIMIC-IV is organized into two main data modules: *hosp* and *icu*. The *hosp* module captures information from the hospital-wide electronic health record (EHR), detailing hospitalizations, patient demographics, laboratory results, medication administration, billing data, and more. In contrast, the *icu* module contains highly detailed clinical data from the ICU, sourced from the MetaVision clinical information system, including treatment plans, and monitoring data (Johnson et al., 2024; Johnson et al., 2023; Goldberger et al., 2000).

Data collection spans from 2008 to 2022, and the records are meticulously de-identified following Health Insurance Portability and Accountability Act (HIPAA) guidelines. Patient identifiers, such as names and social security numbers were removed to protect privacy. Additionally, dates of death are available up to one-year post-discharge and are derived from hospital and state records. Ethical approval for data collection and the creation of this research resource was obtained from the Institutional Review Board at the BIDMC. The dataset was made accessible through a data use agreement, ensuring compliance with all ethical and legal guidelines for data usage. With this agreement in place, the database is available for researchers and academics.

2.2. Variables

The independent variables for this study include socio-demographic, clinical, and resource utilization data. Socio-demographic variables cover gender, age, race, marital status, and type of insurance. Admission characteristics include the type of admission and admission location. Clinical variables included the number of comorbidities, as well as indicators for Hepatitis A, B, C, D, and E, and Hepatic

Coma. Resource utilization variables include counts of medications, procedures, and drugs dispensed before ICU admission. Additional ICU-specific measures capture the volume of fluids prescribed and the number of procedures performed within the ICU. This study aims to predict three primary outcomes: *Total Length of Stay*, measured as the days spent in the ICU or hospital; *Discharge Location*, detailing the destination upon discharge, such as home, hospice, or skilled nursing; and *Discharge Outcome*, a binary variable indicating discharge status (death or alive).

2.3. Data Processing and Preparation

Given the extensive and diverse nature of the MIMIC-IV dataset, specific steps were undertaken using MySQL to refine and tailor the data for this study on hepatitis patients admitted to the ICU (Oracle, 2021). Figure 1 illustrates the data pre-processing workflow, detailing the sequential steps taken to create the final dataset.

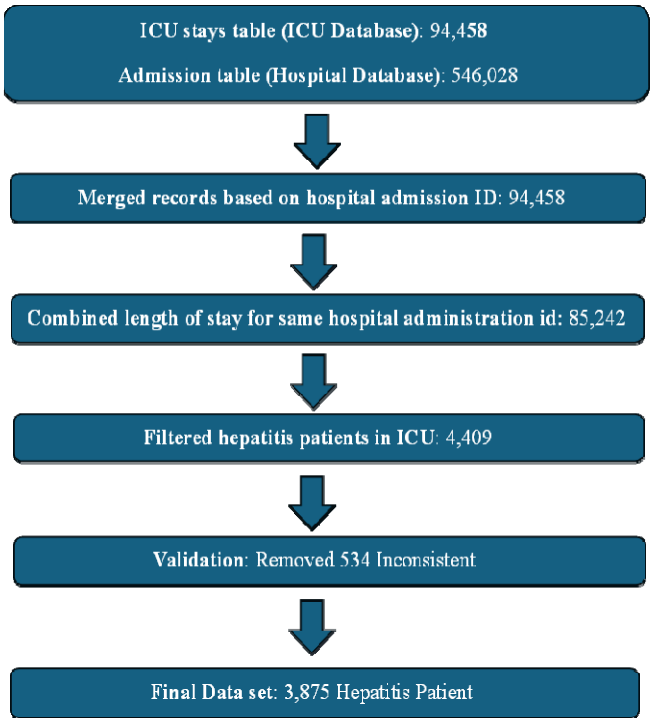


Figure 1. Data preprocessing workflow for extracting and refining hepatitis patient records from the MIMIC-IV database.

a) **International Classification of Diseases (ICD) Code Filtering and Data Integration**: To isolate hepatitis-related hospital admissions, ICD codes specific to viral hepatitis were identified and filtered from a comprehensive table of diagnostic codes. These filtered ICD codes were then linked with a table containing hospital admission identifiers to extract records corresponding to hepatitis diagnoses. Subsequently, the table was merged with another table containing patient demographic information and total length of stay, resulting in a hepatitis-specific patient cohort.

b) **Feature Engineering**: To enrich the dataset, several additional variables were calculated. Pre-ICU and post-ICU medication counts, as well as procedure counts, were derived by aggregating respective records. For patients with multiple hepatitis diagnoses, individual indicator columns for Hepatitis A, B, C, D, and E were created, facilitating detailed analysis of co-infections. A comorbidity count was also computed by excluding hepatitis-related ICD codes and counting the remaining diagnoses.

c) **Data Cleaning and Validation**: The combined dataset initially consisted of 4,409 patient records. A rigorous data validation process identified 534 records with inconsistencies between discharge outcomes and discharge locations. These records were excluded to ensure data integrity, resulting in a final dataset of 3,875 hepatitis patient records. Patients included in this analysis were those admitted to the ICU at Beth Israel Deaconess Medical Center and diagnosed with any form of hepatitis.

d) **Data Transformation**: The data transformation phase involved preparing categorical variables for ML by converting them into factor data types and recategorizing the race variable to address sparsity and improve interpretability. Subcategories with similar characteristics were grouped under broader categories, such as combining all Asian subcategories (e.g., "Asian - Chinese," "Asian - Indian") into "Asian" and all Black subcategories (e.g., "Black/African," "Black/Caribbean Island") into "Black." Hispanic/Latino and White subcategories were similarly merged, while categories with low counts, such as "South American," "Native Hawaiian or Other Pacific Islander," and "Multiple Race/Ethnicity," were consolidated into an "Other" category.

e) **Handling Missing Values:** To address missing values, a thorough check was conducted across all variables, revealing the following counts of missing values per variable: marital status (296), insurance type (64), number of procedures in the ICU (237), number of pre-ICU procedures (642), and medications (248), total length of stay (1), and discharge location (18). Multiple Imputation by Chained Equations (MICE) was employed to impute the missing values (Azur et al., 2011). Predictive Mean Matching (PMM) was used for numeric columns, while Polytomous Logistic Regression (Polyreg) was applied to categorical variables with more than two levels. The imputation process was repeated five times, and the results were pooled to produce a single, complete dataset for analysis.

f) **Handling Imbalanced Data**: The discharge outcome variable, which indicates whether patients were alive or deceased upon discharge, displayed a significant class imbalance, with 3,353 records for the alive category and 522 for death. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was used to balance the classes (Pradipta et al., 2021). This resulted in a revised dataset with 2,088 records labeled as alive and 1,566 as death, ensuring a more equitable representation of both classes for model training.

g) **Training and Test Data**: After data cleaning and preprocessing, the records were partitioned for subsequent modeling tasks. An 80:20 split was used to separate the data into training and testing sets, a crucial step for evaluating model performance.

   I. For the *Discharge Outcome*, after applying SMOTE to address class imbalance, a total of 3,654 records were obtained. Of these, 2,923 records (80%) were used for training, and 731 records (20%) were allocated for testing. The training set was used to train the models, while the testing set was reserved to evaluate predictive performance.

   II. For *Length of Stay* and *Discharge Location*, 3,875 records were available. The same partitioning strategy was applied, with 3,100 records (80%) used for training and 775 records (20%) for testing. This partition ensured that models could be appropriately trained and assessed for their predictive ability across different outcomes.

2.4. Machine Learning Models

For predicting the discharge outcome (death or alive), the discharge location variable was not used as predictor for the analysis. This decision was made because certain discharge location categories overlapped with death records, providing redundant information. The following models were considered:

a) **Model 1: Logistic Regression**: It is a statistical model used for binary classification, estimating the probability of one of two possible outcomes (in this case alive or death) based on a set of predictor variables. This model uses the logistic function to transform predictions into probabilities between 0 and 1 (Stoltzfus, 2011). Logistic Regression was chosen for its interpretability, providing clear insights into how each clinical and socio-demographic feature influences discharge outcome. This characteristic is especially valuable in this research, where

understanding the relative impact of individual predictors on patient survival is crucial for making informed inferences about patient care.

b) **Model 2: Random Forest Classification**: This model was applied to enhance the accuracy of predicting discharge outcome. The model constructs multiple decision trees using random data samples, with each tree voting for a class. The final prediction is determined by the majority vote across all trees. This ensemble approach captures complex, non-linear relationships among variables, offering strong predictive performance and minimizing the risk of overfitting (Wrld, 2024). Although it does not offer the interpretive insights of Logistic Regression, Random Forest's capability to handle heterogeneous data makes it well-suited for maximizing predictive power in clinical scenarios (Couronné et al. 2018).

To predict LoS for ICU-admitted hepatitis patients, we considered two scenarios. In the first case, we considered LoS as a count variable (Fernandez & Vatcheva, 2022), and as a continuous variable (Xu et al., 2022) in our second case, allowing for a comprehensive exploration of different modeling approaches.

a) **Model 1: Generalized Additive Model (GAM) with Negative Binomial (NB) Distribution:** For the discrete case, a GAM with a NB distribution was utilized to address overdispersion, where the variance in the count data exceeds the mean (Zeileis et al., 2015). By incorporating smooth functions, the GAM framework allowed for modeling potential non-linear relationships between predictors and LoS. This approach provided the necessary flexibility to capture complex patterns in the data, ensuring a robust and accurate analysis (Hastie & Tibshirani, 2014).

b) **Model 2: Random Forest Regression:** For the continuous case, Random Forest Regression was employed. This ensemble method constructs multiple decision trees and averages their predictions to capture non-linear relationships and interactions among predictors (Wrld, 2024). Using Random Forest provided an alternative view, enabling a performance comparison between the count-based and continuous modeling approaches.

To predict the discharge location of ICU-admitted hepatitis patients, models were trained after excluding discharge outcome variable which is not relevant to this analysis.

a) **Model 1: Gradient Boosting Model (GBM):** Gradient Boosting is an ensemble learning technique that builds multiple decision trees sequentially, with each new tree correcting errors made by the previous ones. This model focuses on reducing bias and improving predictive performance through boosting. GBM was chosen for its strong ability to handle complex, non-linear relationships and provide high accuracy (Zhang et al. 2019), which is critical in predicting discharge locations, a multi-class categorical outcome.

b) **Model 2: Multinomial Logistic Regression**: Multinomial Logistic Regression is an extension of logistic regression used when the dependent variable has more than two categories (Kwak & Clayton-Matthews, 2002). It models the probability of each class as a function of the predictors, providing a straightforward and interpretable approach to multi-class classification. This model was included to establish baseline and offer a simpler, interpretable model for understanding the impact of predictor variables on different discharge locations.

2.5. Model Evaluation

The models were evaluated using appropriate performance metrics based on the type of outcome variable:

- **Regression Models:** For LoS, regression models were evaluated using Root Mean Squared Error (RMSE), R-squared (R^2), and Mean Absolute Error (MAE). RMSE measures the average magnitude of error between predicted and actual values, penalizing larger deviations more heavily, making it particularly sensitive to outliers. R^2 quantifies the proportion of variation in the dependent variable explained by the model, offering an overall measure of fit. MAE, unlike

RMSE, calculates the average of the absolute errors, providing an intuitive measure of model accuracy by treating all errors equally (Zwanenburg, 2022).

- **Multiclass Classification Models:** These were assessed using Accuracy, the Kappa statistic, the Brier Score, and the Area Under the Receiver Operating Characteristic Curve (ROC AUC). Accuracy reflects the percentage of correctly classified instances, while the Kappa statistic accounts for agreement by chance, providing a more robust evaluation of model performance. The Brier Score measures the mean squared difference between predicted probabilities and the actual class, with lower scores indicating better calibration. ROC AUC quantifies a model's ability to distinguish between classes, with values closer to 1 denoting superior performance (Zwanenburg, 2022).
- **Binary Classification Models:** For binary outcomes, Logistic Regression and Random Forest Classification models were evaluated using Brier Score, Accuracy, Kappa, ROC AUC, Sensitivity, and Specificity. Sensitivity, also known as recall, indicates the proportion of true positives correctly identified by the model, while Specificity measures the proportion of true negatives correctly identified. These metrics provide insights into the model's ability to balance false positives and false negatives. The Brier Score evaluates the accuracy of probabilistic predictions, with lower values reflecting better performance. ROC AUC measures the discriminatory power of the model across thresholds, while Accuracy and Kappa offer overall measures of correctness, with Kappa adjusting for chance agreement (Zwanenburg, 2022).

2.6. Resampling Techniques

To ensure the robustness and generalizability of the models, 10-fold cross-validation was employed. This approach divides the dataset into 10 subsets (or folds). The model is trained on 9 folds and validated on the remaining fold in each iteration. This process is repeated 10 times, with each fold serving as the validation set once, and the average performance across all iterations is reported. The choice of K=10 is supported by experimental analysis and observations, as it provides a balance between computational efficiency and reliability (Verma et al., 2024). K-fold cross-validation is critical for mitigating the risk of overfitting, as it ensures that the model is tested on diverse subsets of the data, not just a single train-test split. This approach provides a more reliable estimate of the model's performance on unseen data by leveraging all data points for both training and validation at different stages. Particularly for complex datasets, such as the one used in this study, 10-fold cross-validation enhances the reliability and generalizability of the results, making it a robust validation strategy for machine learning applications (Wilimitis & Walsh, 2023).

2.7. Software

Data analysis was conducted in RStudio (R Core Team, 2024; RStudio Team, 2022) using various packages to streamline data manipulation, model development, and visualization. MySQL (Oracle, 2024) was used to extract and manage data from the MIMIC-IV database. In RStudio, data processing was performed using the *tidyverse* (Wickham et al., 2019), while missing data were imputed with *mice* (Buuren & Groothuis-Oudshoorn, 2011), and class imbalance was addressed using *DMwR* (Torgo, 2024). Model development and evaluation were facilitated by MachineShop (Smith, 2024), enabling a streamlined approach to implementing and comparing various predictive models. Statistical modeling was conducted using *MASS* (Venables & Ripley, 2002), which supported the fitting of Negative Binomial regression models, and *mgcv* (Wood, 2011), which was utilized for constructing Generalized Additive Models (GAMs) to capture nonlinear relationships. Random Forest modeling was implemented using the *randomForest* package (Liaw & Wiener, 2002). Data visualizations were created with *ggplot2* (Wickham, 2016), while descriptive statistics and baseline characteristics were summarized using *tableone* (Yoshida & Bartel, 2022).

3. RESULTS

3.1. Demographic and Descriptive Statistics

The study included a total of 3,875 patients admitted to the ICU with hepatitis-related conditions. The gender distribution was predominantly male, accounting for 68.2% (n = 2,644) of the sample, while females constituted 31.8% (n = 1,231). The average age of the participants was 53.26 years (SD = 12.77). The racial composition was primarily White (57.3%, n = 2,220), followed by Black (16.9%, n = 653), Hispanic/Latino (6.3%, n = 245), and Asian (6.0%, n = 234). Other racial groups, such as Indian/Alaska Native and Portuguese, made up a smaller proportion of the population, each representing less than 1% of the sample. These demographic characteristics, along with details on marital status, insurance type, clinical and admission characteristics, are summarized in **Table 1**. The distribution of hepatitis types is shown in **Figure 2**, which reveals that Hepatitis C was the most prevalent condition, affecting 82.2% (n = 3,186) of the total sample. Hepatitis B was the second most common condition at 19.3% (n = 746), while Hepatitis A (1.0%, n = 37), Hepatitis D (9.2%, n = 357), and Hepatitis E (0.1%, n = 3) were less frequent. Additionally, 8.0% (n = 311) of patients experienced hepatic coma.

Table 1. Baseline demographic, clinical, and admission characteristics of study participants

N	3,875
GENDER = Male (%)	2644 (68.2)
AGE (MEAN (SD))	53.26 (12.77)
RACE (%)	
INDIAN/ALASKA NATIVE	10 (0.3)
ASIAN	234 (6.0)
BLACK	653 (16.9)
HISPANIC/LATINO	245 (6.3)
OTHER	120 (3.1)
PORTUGUESE	29 (0.7)
UNKNOWN	364 (9.4)
WHITE	2220 (57.3)
MARITAL STATUS (%)	
DIVORCED	398 (10.3)
MARRIED	1106 (28.5)
NA	296 (7.6)
SINGLE	1873 (48.3)
WIDOWED	202 (5.2)
INSURANCE (%)	
MEDICAID	1542 (39.8)
MEDICARE	1472 (38.0)
NA	64 (1.7)
OTHER	123 (3.2)
PRIVATE	674 (17.4)
TYPE OF ADMISSION (%)	
DIRECT EMERGENCY	230 (5.9)
DIRECT OBSERVATION	10 (0.3)
ELECTIVE	50 (1.3)

EU OBSERVATION	5 (0.1)
EMERGENCY WARD	2296 (59.3)
OBSERVATION ADMIT	548 (14.1)
SURGICAL SAME DAY ADMISSION	155 (4.0)
URGENT	581 (15.0)
ADMISSION LOCATION (%)	
EMERGENCY ADMISSIONS	1998 (51.6)
INTERNAL TRANSFERS	45 (1.2)
REFERRALS	927 (23.9)
TRANSFER FROM OTHER FACILITY	898 (23.2)
UNKNOWN	7 (0.2)
No. of COMORBIDITIES (MEDIAN [IQR])	18.00 [12.00, 24.00]
HEPATITIS_A = 1 (%)	37 (1.0)
HEPATITIS_B = 1 (%)	746 (19.3)
HEPATITIS_C = 1 (%)	3186 (82.2)
HEPATITIS_D = 1 (%)	357 (9.2)
HEPATITIS_E = 1 (%)	3 (0.1)
HEPATIC_COMA = 1 (%)	311 (8.0)
PRE-ICU MEDICATIONS COUNT (MEDIAN [IQR])	82.00 [36.00, 140.25]
PRE-ICU PROCEDURES COUNT (MEDIAN [IQR])	3.00 [2.00, 6.00]
DISPENSED MEDICATIONS COUNT (MEDIAN [IQR])	65.50 [33.00, 118.50]
ICU MEDICATIONS COUNT (MEDIAN [IQR])	54.00 [20.00, 154.00]
ICU PROCEDURES COUNT (MEDIAN [IQR])	7.00 [3.00, 14.00]
TOTAL LENGTH OF STAY (MEDIAN [IQR])	2.00 [1.00, 4.00]
DISCHARGE_LOCATION (%)	
AGAINST MEDICAL ADVICE	170 (4.4)
DEATH	522 (13.5)
HOME-BASED CARE	1997 (51.5)
HOSPICE CARE	100 (2.6)
NA	18 (0.5)
OTHER HEALTHCARE FACILITY	110 (2.8)
PSYCHIATRIC CARE	83 (2.1)
REHABILITATION	191 (4.9)
SKILLED NURSING/LONG-TERM CARE	684 (17.7)
DISCHARGE OUTCOME = 1 (%)	522 (13.5)

The distribution of hepatitis types across racial groups is depicted in **Figure 3**. Hepatitis C was the most prevalent condition across all racial groups, with White patients contributing the largest proportion. Among Black and Hispanic/Latino patients, Hepatitis C was also the most common condition but showed slightly smaller proportions compared to White patients, while Hepatitis B was notably more prevalent among Asian patients compared to other racial groups. In addition, Hepatitis B disproportionately affects non-

White individuals compared to Whites. Hepatitis A, D, and E were observed in smaller proportions across all racial groups, with Hepatitis E being particularly rare. Hepatic Coma was most common among White patients, with smaller contributions from other racial groups.

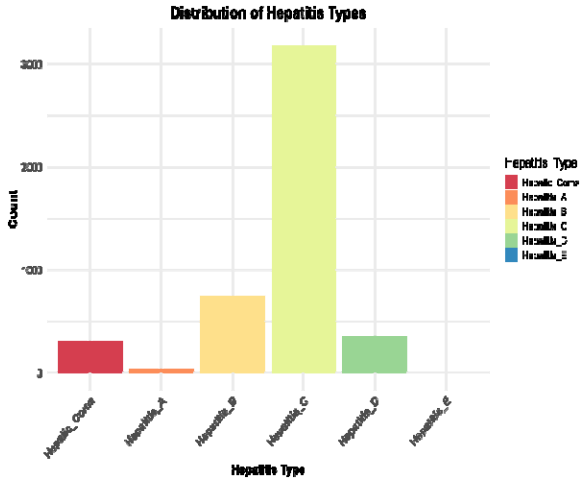


Figure 2. Distribution of hepatitis types among study participants

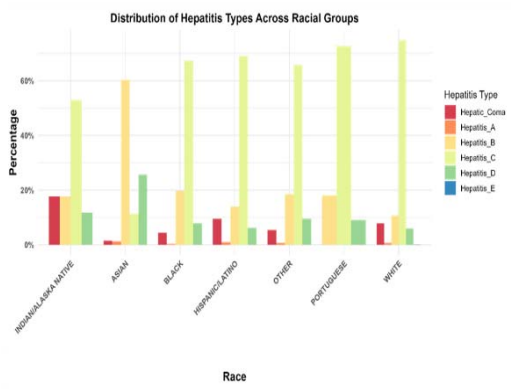


Figure 3. Distribution of hepatitis across racial groups

The distribution of total length of stay for hepatitis patients was highly skewed as commonly seen in hospital records, with most patients having a short hospital stay (**Figure 4**). The median total length of stay was 2 days (IQR: 1–4 days) as stated in **Table 1**, with some patients having an extended stay of over 30 days. Variations in length of stay were also observed based on admission location (**Figure 5**). For instance, patients admitted from emergency admissions had shorter median stays compared to those transferred from other facilities. Notably, patients transferred from other facilities exhibited a wider range of lengths of stay, including outliers with significantly extended hospitalizations. These trends highlight how admission pathways may influence the duration of ICU stays.

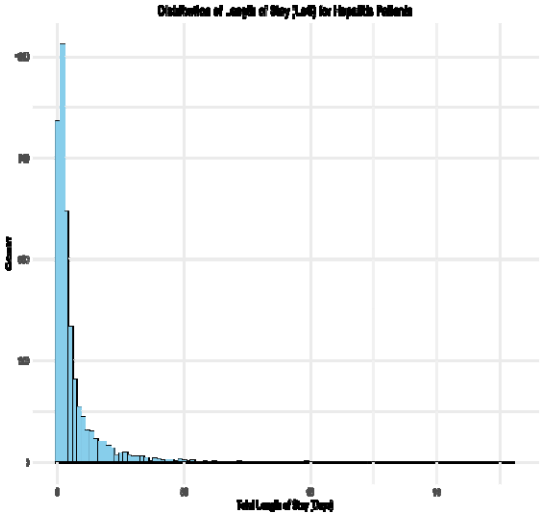


Figure 4. Distribution of total LoS for hepatitis patients

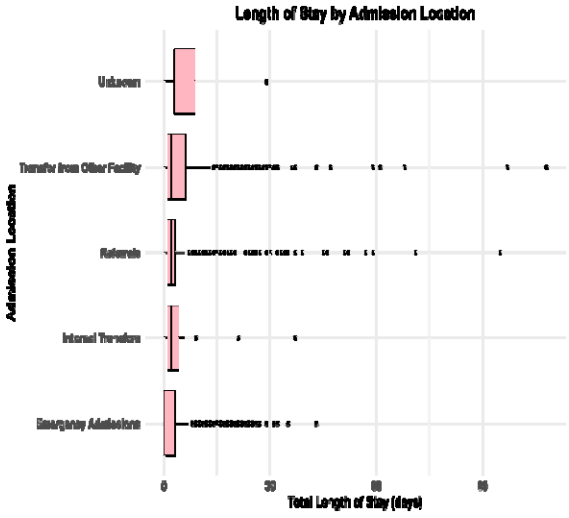


Figure 5. Length of stay by admission location

The boxplot in **Figure 6** illustrates the relationship between the total length of stay and discharge outcome. Patients with longer lengths of stay were more likely to have died, as indicated by the higher median and wider range of LoS in the "Died" group compared to the "Survived" group. **Figure 7** highlights the distribution of discharge locations among patients. The majority (51.5%, n = 1,997) were discharged to home-based care. Skilled nursing or long-term care facilities accounted for 17.7% (n = 684) of discharges, while 13.5% (n = 522) of patients died. Other notable discharge destinations included rehabilitation centers (4.9%, n = 191), hospice care (2.6%, n = 100), psychiatric care (2.1%, n = 83), and other healthcare facilities (2.8%, n = 110). A small proportion of patients left the hospital against medical advice (4.4%, n = 170).

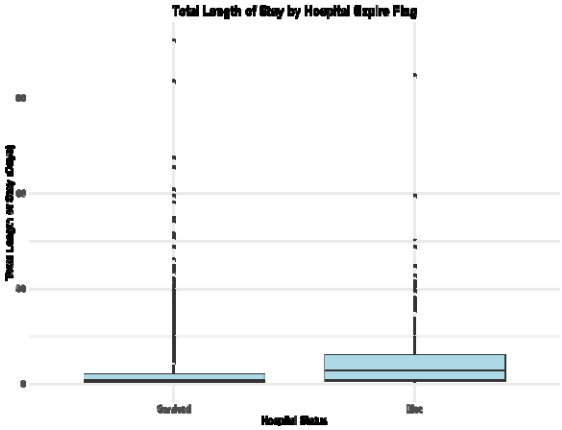


Figure 6. Discharge outcome based on total LoS

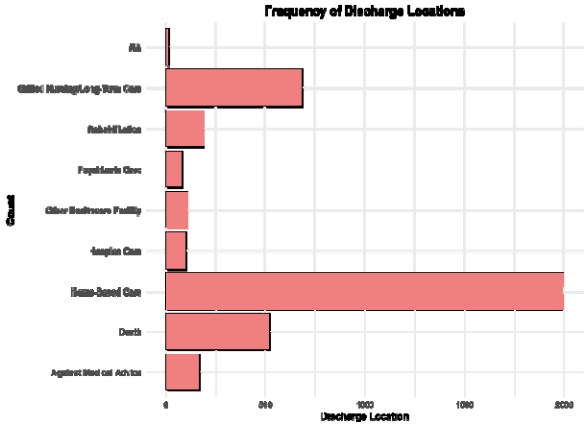


Figure 7. Distribution of discharge locations among patients

3.2. Model Results and Performance

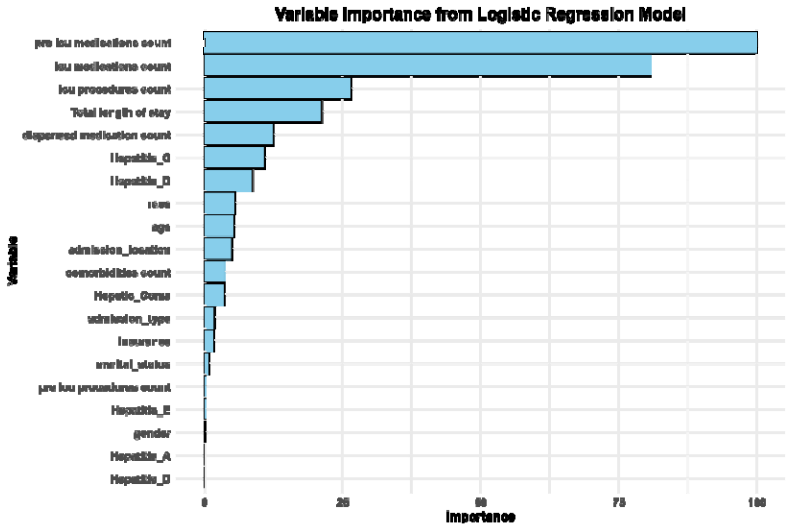
Discharge Outcome (Died/Alive)

The Logistic Regression and Random Forest models were developed and evaluated to predict discharge outcomes (died/alive) for ICU-admitted hepatitis patients. These models were tested on test data and through cross-validation. Performance metrics for both models are presented in **Table 2**. On the test data, Logistic Regression yielded a Brier Score of 0.1302, an accuracy of 0.818, a Kappa of 0.622, and an ROC AUC of 0.887. Sensitivity and specificity were 0.73 and 0.87, respectively. The Random Forest model demonstrated a Brier Score of 0.0875, an accuracy of 0.87, a Kappa of 0.742, and an ROC AUC of 0.95. Sensitivity and specificity for the Random Forest model were 0.78 and 0.94, respectively. Cross-validation results are also shown in **Table 2**, indicating the mean and standard deviation for each metric across repeated samples. The mean Brier Score for Random Forest was 0.089 (±0.010), while Logistic Regression had a mean Brier Score of 0.129 (±0.011). The Random Forest model had a mean accuracy of 0.87 (±0.017) and a mean Kappa of 0.729 (±0.037), compared to Logistic Regression's mean accuracy of 0.82 (±0.02) and mean Kappa of 0.629 (±0.043). Moreover, statistical comparisons were conducted to evaluate differences in performance metrics between the Logistic Regression and Random Forest models. The results indicated statistically significant differences ($p < 0.001$) for all key metrics, including Brier Score, accuracy, Kappa, ROC AUC, sensitivity, and specificity. These findings demonstrated that the Random Forest model consistently outperformed Logistic Regression across all evaluated criteria. Figures 8 and 9 display the variable importance for the Logistic Regression and Random Forest models, respectively. The top predictors in the Logistic Regression model **(Figure 8)** were pre-ICU medications count, ICU medications count, and ICU procedures count, with sociodemographic factors such as race

and age also ranking among the top 10 variables. For the Random Forest model **(Figure 9)**, ICU medications count, ICU procedures count, and total LoS were identified as the most influential variables, with race and age again appearing among the top 10 predictors. These findings highlight the consistent importance of race and age as predictors, emphasizing the need to consider both clinical and sociodemographic factors when analyzing discharge outcomes.

Table 2. *Performance metrics of Logistic Regression and Random Forest models on test and cross-validated data*

Metrics	Logistic Regression		Random Forest	
	Test data	Cross-validation	Test data	Cross-validation
Brier Score	0.1301607	0.129 ± 0.011	0.08772361	0.089 ± 0.010
Accuracy	0.8166895	0.821 ± 0.020	0.87551300	0.870 ± 0.017
Kappa	0.6220590	0.629 ± 0.043	0.74208369	0.729 ± 0.037
ROC AUC	0.8865039	0.889 ± 0.017	0.94999924	0.948 ± 0.012
Sensitivity	0.7350158	0.736 ± 0.040	0.78233438	0.780 ± 0.033
Specificity	0.8792271	0.884 ± 0.019	0.94685990	0.937 ± 0.016



Figures 8: Variable importance for Logistic Regression

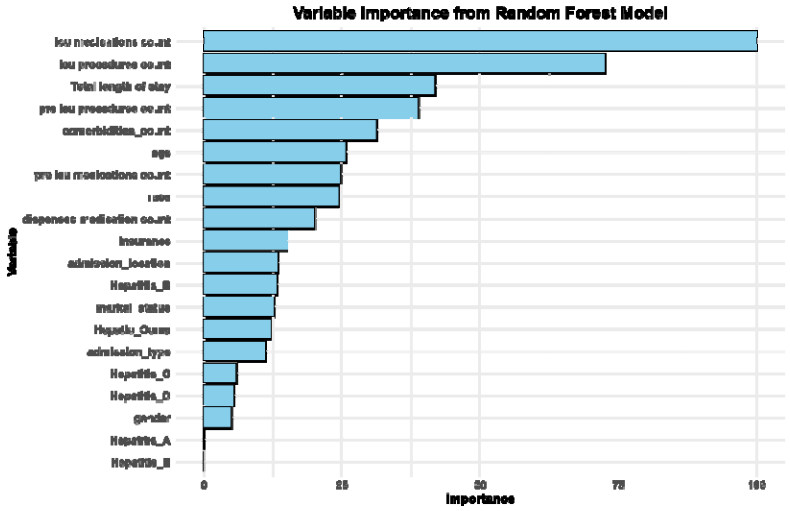
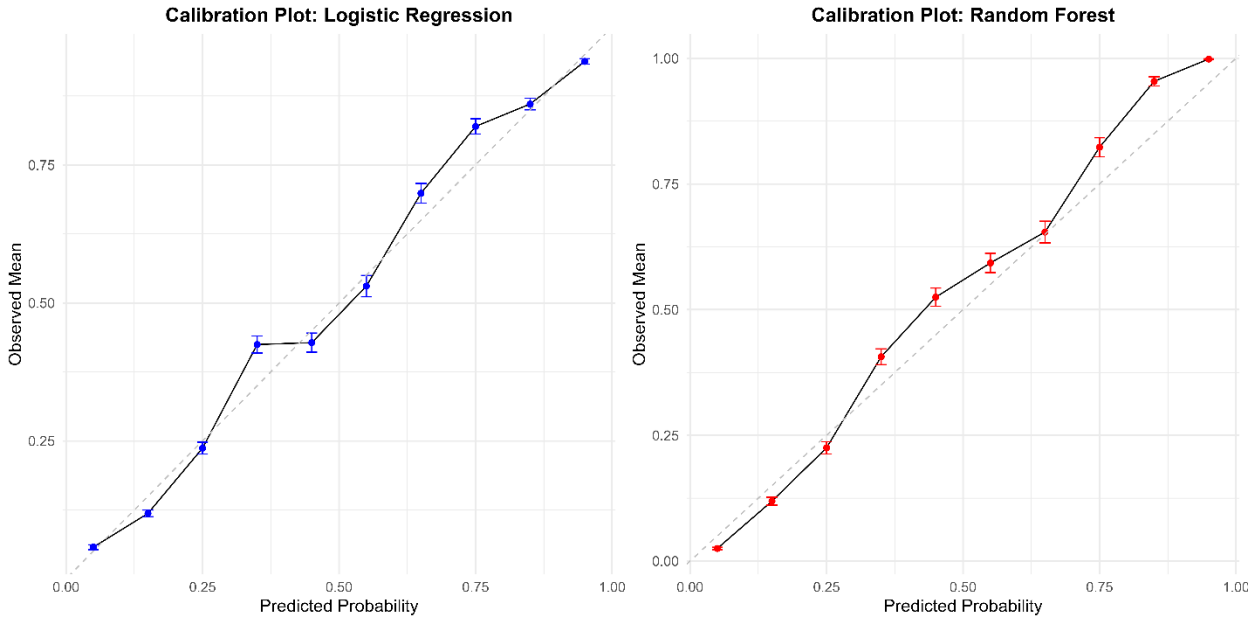


Figure 9. Variable importance for Random Forest classification

The calibration curves shown in **Figure 10** depict the relationship between predicted probabilities and observed outcomes. **Figure 10a** illustrates the calibration curve for the Logistic Regression model, which showed reasonable alignment, with minor deviations at the extremes. While **Figure 10b** represents the calibration curve for the Random Forest model, which demonstrated closer alignment to the diagonal line, indicating better calibration.



(a) **(b)**

Figure 10. Calibration curves: (a) Logistic Regression, (b) Random Forest.

Table 3 summarizes the confusion matrix results for both models. The Logistic Regression model showed an agreement of 0.443 for observed class 0 and 0.2974 for class 1, with sensitivity and specificity values of 0.773 and 0.696, respectively. The Random Forest model had higher agreement, with 0.483 for class 0 and 0.326 for class 1, a sensitivity of 0.844, and a specificity of 0.763.

Table 3. Confusion matrix metrics for Logistic Regression and Random Forest models

Metric	Logistic Regression		Random Forest	
	Class 0	Class 1	Class 0	Class 1
Observed	0.5726993	0.4273007	0.5726993	0.4273007
Predicted	0.5729394	0.4270606	0.5847769	0.4152231
Agreement	0.4431088	0.2974701	0.4836014	0.326125
Sensitivity	0.7737198	0.6961610	0.8444246	0.7632218
Specificity	0.6961610	0.7737198	0.7632218	0.8444246
PPV (Precision)	0.7733956	0.6965524	0.8269844	0.7854216
NPV	0.6965524	0.7733956	0.7854216	0.8269844

Figures 11 and **12** illustrate the confusion matrix results for the **Logistic Regression** and **Random Forest** models, respectively. In **Figure 12**, the Random Forest model demonstrates a stronger alignment along the diagonal, indicating higher agreement between predicted and observed classes. The color gradients in both **Figures 11 and 12** represent the predicted probabilities, with darker shades reflecting higher probabilities of correct classification. Overall, the Random Forest model exhibits improved

sensitivity and specificity compared to the Logistic Regression model, aligning with the numerical results presented in **Table 3**.

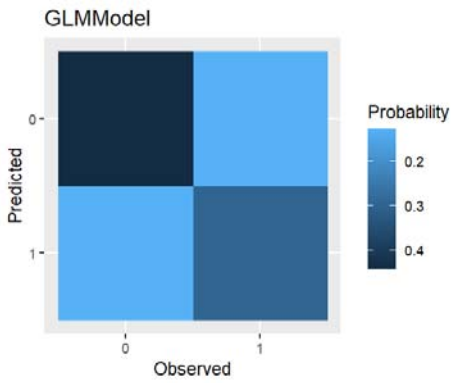


Figure 11. Confusion matrix for Logistic Regression

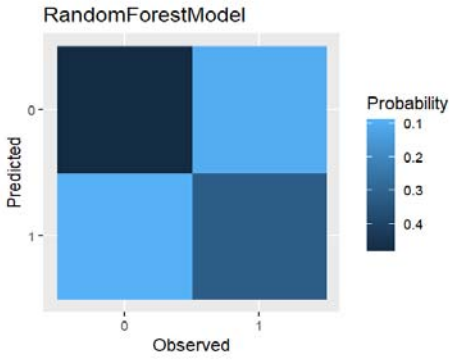


Figure 12. Confusion matrix for Random Forest

Total Length of Stay

While the previous models focused on classification outcome, the Generalized Additive Model (GAM) with Negative Binomial distribution and the Random Forest Regression model were utilized to predict total length of stay as discrete and continuous outcome respectively in the test data and through cross-validation. The Generalized Additive Model (GAM) with a Negative Binomial family was applied to predict LoS, as the mean in the dataset for LoS was 3.69, with a variance of 42.54, indicating overdispersion. On the test data, the model achieved a Root Mean Squared Error (RMSE) of 2.9619, a Mean Absolute Error (MAE) of 1.4237, and an R-squared (R^2) of 0.7594, demonstrating its ability to explain 75.94% of the variance in the LoS (**Table 4**). The deviance for the model was 2772.78, and the Akaike Information Criterion (AIC) was 10410.0, reflecting the model's goodness-of-fit and complexity. Cross-validation was conducted with 10 folds to assess the model's robustness. The cross-validated RMSE was 5.5212, while the mean R-squared (R^2) was 0.7601, and the mean MAE was 1.5634. These cross-validation results highlight slightly higher variability and prediction errors compared to the test data, likely due to the presence of extreme values or overdispersion in the data. Furthermore, as shown in **Figure 13**, the majority of predictions are closely aligned with actual values, deviations are observed at higher LoS, indicating potential challenges in predicting extended stays.

Table 4. Performance metrics of General Additive Model with Negative Binomial

Metric	Test Data	Cross-Validation
Root Mean Squared Error	2.9619	5.5212
Mean Absolute Error	1.4237	1.5634
R-squared	0.7594	0.7601

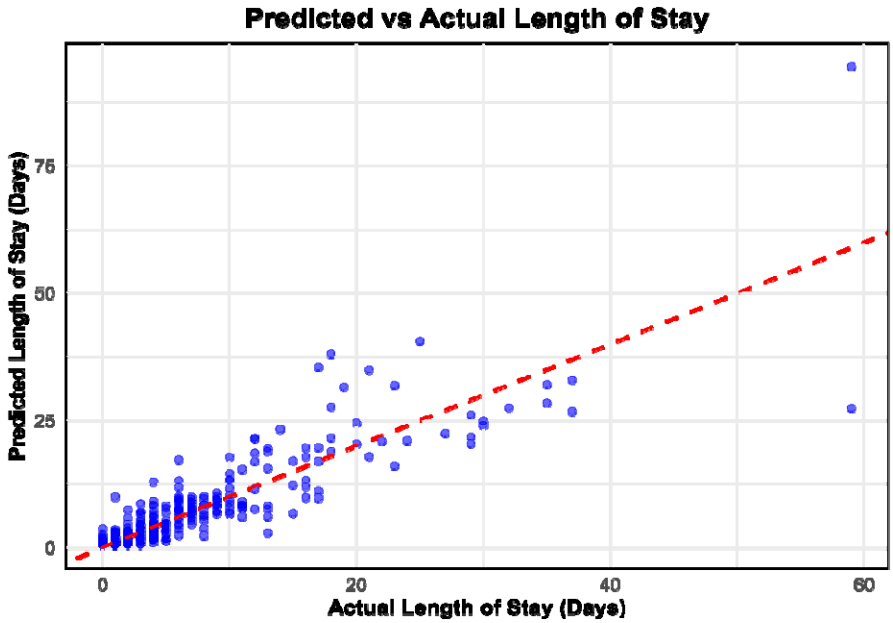


Figure 13: Predicted vs. Actual Length of Stay for the Generalized Additive Model (GAM)

Table 5. Performance Metrics of Random Forest Regression Models

Metric	Test Data	Log Transformed Test Data	Cross-Validation
Root Mean Squared Error	3.271	0.3144	0.295 (±0.020)
Mean Absolute Error	1.360	0.2377	0.229 (±0.014)
R-squared	0.755	0.821	0.838 (±0.021)

The Random Forest model was subsequently applied after log-transforming the total Length of Stay (LoS), which resulted in improved performance, achieving an RMSE of 0.3144, an R-squared (R^2) of 0.821, and an MAE of 0.2377 on the test data. Cross-validation further confirmed the model's consistency, with a mean RMSE of 0.295 (±0.020), a mean R^2 of 0.838 (±0.021), and a mean MAE of 0.229 (±0.014). However, when back transformed to the original scale, the performance metrics were: RMSE of 3.271, R^2 of 0.755, and MAE of 1.360 (**Table 5**).

Figure 14 illustrates the variable importance of the Random Forest Regression model for predicting length of stay. The most influential predictors were the number of ICU medications and procedures, comorbidities count, and pre-ICU procedure count. While clinical factors were most prominent, sociodemographic variables such as age and race also emerged as notable predictors, highlighting their potential impact on length of stay alongside clinical characteristics. The scatter plots in **Figure 15** illustrate the relationship between predicted and actual Length of Stay values for the Random Forest Regression model. **Figure 15a** presents the scatter plot for the log-transformed data, where a strong linear relationship is observed, with predictions closely aligning with the actual values. In contrast, **Figure 15b** displays the scatter plot for the original Length of Stay data, where greater dispersion is evident, particularly for higher values.

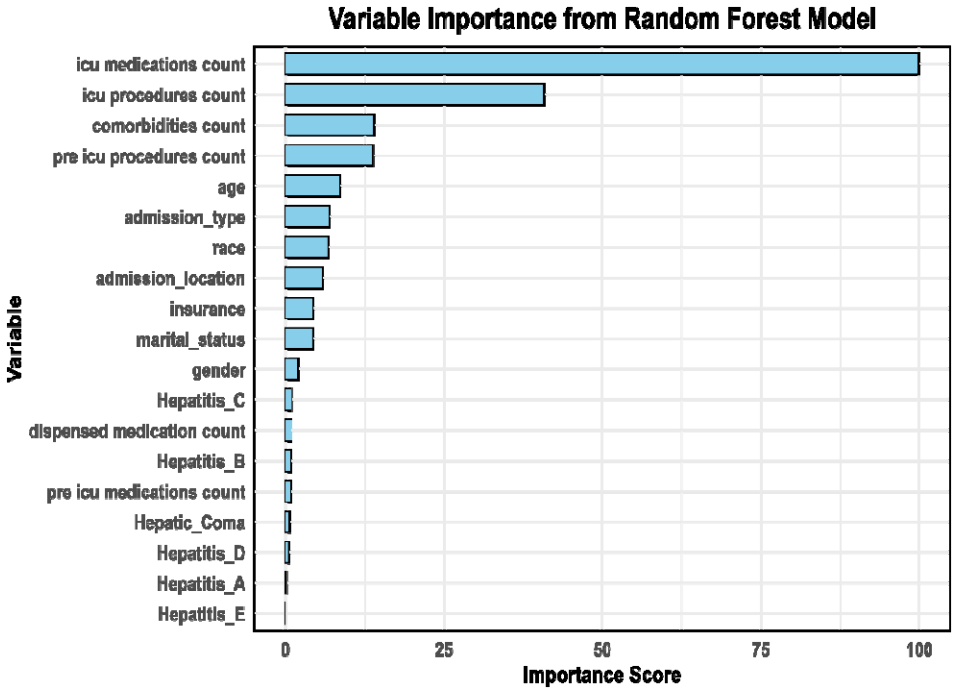
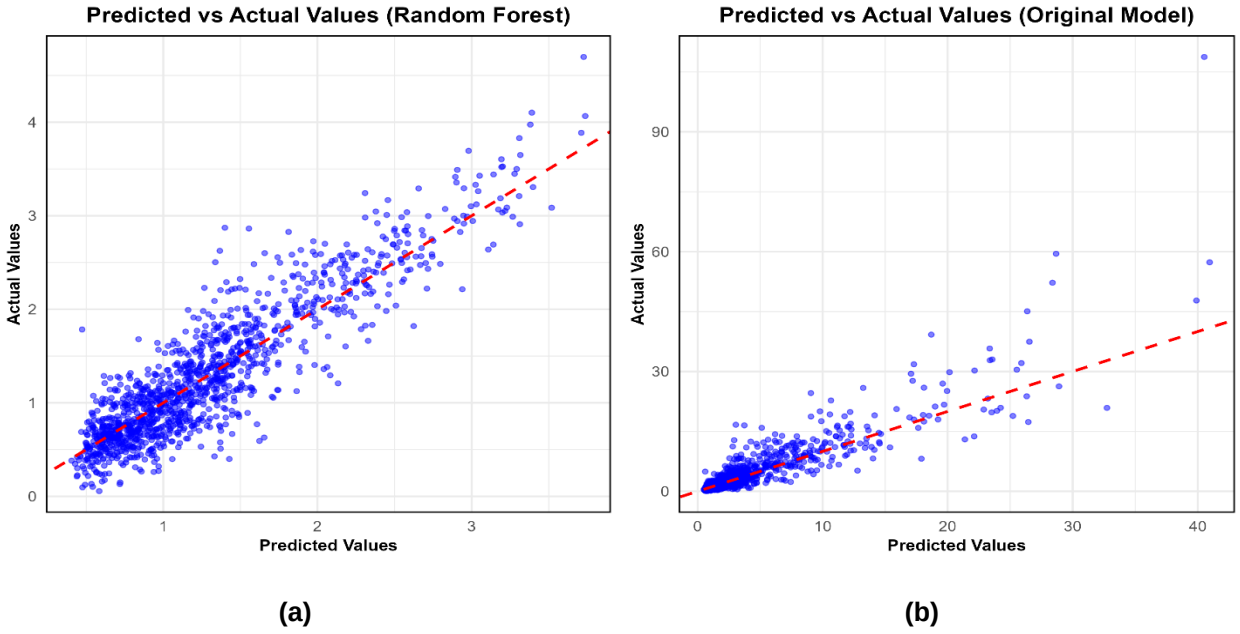


Figure 14. Variable Importance for Random Forest Regression Model



(a) **(b)**

Figure 15: Predicted vs. Actual Length of Stay for the Random Forest Regression Model: (a) Log-Transformed Data and (b) Normal Data.

Discharge Location

Building on the evaluation of model performance for previous outcomes, the focus now turns to predicting discharge location. The Gradient Boosting Model (GBM) was evaluated on test data, achieving a Brier Score of 0.589, an accuracy of 0.558, a Kappa of 0.2110, and an ROC AUC of 0.740 (**Table 6**). Cross-validation analysis showed a mean Brier Score of 0.5950 (±0.0206), a mean accuracy of 0.559 (±0.227), and a mean Kappa of 0.2077 (±0.0391), indicating stable performance across folds. The Multinomial Regression Model produced an accuracy of 0.5639 (95% CI: 52.81% - 59.91%) on the test data, with a Kappa value of 0.2037 (**Table 6**). The decay tuning parameter during cross-validation yielded consistent results, with accuracy values around 0.5658 (±0.153) and Kappa values near 0.2335 (±0.30). Significant predictors ($p < 0.05$) identified in this model included gender, marital status, insurance type, type of admission, number of comorbidities, and number of ICU procedures and medications. **Figure 16** illustrates the variable importance of the Gradient Boosting Model, highlighting the number of ICU medications, comorbidities, age, total length of stay and race as the most influential predictors.

Table 6. Performance metrics of Gradient Boosting and Multinominal Regression model on test and cross-validated data

Metric	Gradient Boosting Model		Multinominal Regression	
	Test Data	Cross Validation	Test Data	Cross Validation
Accuracy	55.87%	0.5591 (±0.227)	0.5639	0.5658 (±0.153)
Kappa	0.2110	0.2077 (±0.0391)	0.2037	0.2335 (±0.30)
Brier Score	0.5898	0.5950 (±0.0206)	-	-
ROC AUC	0.7407	0.7319 (±0.0282)	-	-

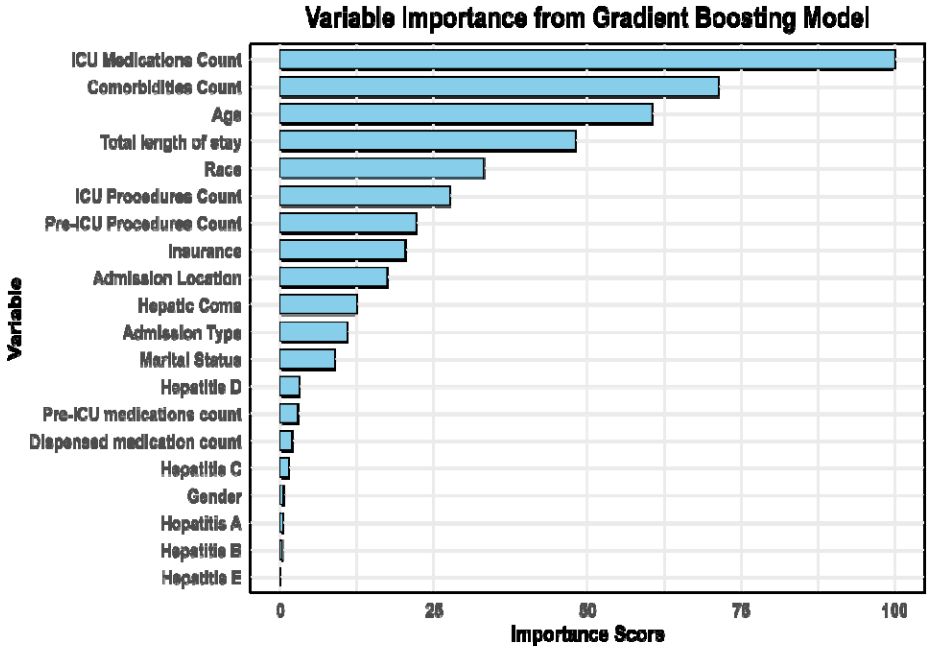


Figure 16. Variable importance for Gradient Boost model

4. DISCUSSION

Despite significant advancements in critical care for hepatitis-related conditions, this study highlights the complexities associated with ICU LoS, discharge outcomes (died/alive), and discharge locations. Leveraging ML models, we successfully predicted these outcomes and identified key predictors, including ICU procedures, medication counts, comorbidities, and sociodemographic factors such as age and race. Notably, race consistently emerged as one of the top predictors across all models, underscoring its critical role in influencing health outcomes for hepatitis patients. This finding is particularly significant as it aligns with observed racial disparities in the raw data, suggesting systemic inequities. Additionally, Hepatitis C emerged as the most prevalent condition in our cohort, aligning with global epidemiological trends that underscore its significant burden on critical care systems (Brunner & Bruggmann, 2021; Petruzziello et al., 2016). These findings emphasize the potential of predictive analytics in not only identifying key drivers of health outcomes but also in addressing disparities, optimizing resource allocation, improving patient outcomes, and reducing healthcare costs.

This study investigated the prediction of ICU LoS for hepatitis patients using ML models, including the Generalized Additive Model (GAM) with a Negative Binomial distribution and Random Forest regression. Random Forest demonstrated superior predictive accuracy, particularly after applying a log transformation to stabilize variance. The log-transformed Random Forest model achieved an R^2 of 0.821 and an RMSE of 0.3144 on test data, effectively capturing complex interactions among predictors. By comparison, GAM with a Negative Binomial distribution achieved an R^2 of 0.7594 and RMSE of 2.9619, effectively addressing overdispersion but yielding higher errors. The key predictors for LoS included ICU medication counts, ICU procedure counts, comorbidities, pre-ICU procedures, and patient age. These findings align with prior studies, such as Gwynn et al. (2019), which demonstrated that higher procedural and pharmacological interventions often correlate with disease severity and resource intensity, leading to longer ICU stays. Similarly, Tola Getachew Bekele et al. (2024) emphasized that prolonged ICU stays are associated with complications, readmissions, and sedative use, underscoring the importance of clinical management strategies to optimize LoS. Although the models performed well for shorter stays, predicting longer LoS posed challenges due to increased variability and extreme values, as reflected in scatter plot (**Figures 13**). Xu et al. (2022) observed similar challenges, reporting that ML models often struggle to predict longer stays due to the inherent variability of extended ICU durations. This study reflects these findings, with greater error dispersion observed in the upper range of LoS predictions. The performance of Random Forest in predicting LoS aligns with Iwase et al. (2022), who reported high predictive accuracy for ICU stays using Random Forest. Ensemble methods such as Random Forest excel in capturing nonlinear interactions and handling heterogeneous predictors, as further evidenced by Alghatani et al. (2021), who achieved 65% accuracy for LoS predictions in ICU datasets. While GAM with Negative Binomial distribution effectively addressed overdispersion, its higher cross-validation errors indicate that variability in extended LoS remains a challenge. Future research could incorporate additional clinical variables or hierarchical modeling techniques to enhance predictions for longer LoS.

The prediction of discharge outcomes (alive vs. death) was conducted using Logistic Regression and Random Forest models. Random Forest outperformed Logistic Regression across all metrics, achieving an accuracy of 87.5%, a Kappa of 0.742, and an ROC AUC of 0.95 on test data. Logistic Regression achieved slightly lower metrics, with an accuracy of 81.67%, a Kappa of 0.622, and an ROC AUC of 0.89. These results highlight the robustness of ensemble methods like Random Forest in handling complex interactions among clinical and demographic predictors. Statistically significant predictors for discharge outcomes included ICU medication counts, ICU procedure counts, and total LoS, all of which reflect disease severity and the intensity of care and key variables in predicting LoS. Prolonged LoS has been linked to increased mortality, consistent with Moitra et al. (2016), who found that each additional ICU Day beyond seven days significantly raises mortality risk. Similar findings were reported by Lingsma et al. (2018) and Tola Getachew Bekele et al. (2024), who highlighted the correlation between extended LoS and mortality, emphasizing its role as both a predictor and consequence of severe illness.

Sociodemographic factors also played a crucial role in predicting discharge outcomes. Race, for instance, emerged as a significant predictor in our study, aligning with Olanipekun et al. (2021), who found disparities in ICU mortality among demographic groups. Age was another critical variable, consistent with Saadatmand et al. (2022), who identified age and Lactate dehydrogenase (LDH) levels as significant determinants of mortality. While LDH levels were unavailable in our dataset, the inclusion of demographic and clinical predictors supported robust outcome predictions.

The prediction of discharge location presented unique challenges, primarily due to the small dataset of 3,875 records and eight discharge categories. Gradient Boosting Model (GBM) and Multinomial Regression yielded moderate performance, with accuracies of 55.87% and 56.39% and Kappa values of 0.2110 and 0.2037, respectively. These results highlight the complexities of multi-class prediction in critical care settings with limited data. The Synthetic Minority Oversampling Technique (SMOTE) was applied to address class imbalance but resulted in data loss, which limited the model's performance. Abad et al. (2021) addressed similar challenges using hierarchical classifiers and SMOTE to manage imbalanced multi-class datasets, suggesting that advanced techniques and larger datasets are critical for improving predictive accuracy. Additionally, as noted by Abad et al. (2021), discharge decisions are influenced by subjective factors such as caregiver preferences, resource availability, and social support, which may not be fully captured by clinical and demographic data alone. Top predictors for discharge location included gender, marital status, insurance type, admission type, number of comorbidities, and ICU procedures and medications. These findings align with Mickle and Deb (2022), who highlighted the importance of clinical and patient characteristics in discharge planning. The association of ICU procedures and medications with discharge location reflects the intensity of care required, further emphasizing their relevance in critical care contexts. Future studies should explore hierarchical classifiers, minimize the number of discharge categories, and incorporate additional variables such as functional status and physiological scores to enhance predictive performance. As highlighted by Mickle and Deb (2022), integrating larger datasets and advanced modeling techniques can provide more clinically actionable insights.

This study has several limitations that warrant further consideration. First, the reliance on the MIMIC-IV database, which reflects a single institution's patient population and care practices, may limit the generalization of the findings to other healthcare settings with differing demographics or protocols. Second, the imbalanced distribution of outcomes, particularly for rare discharge destinations (e.g., hospice, against medical advice), may have reduced model stability and predictive accuracy. While ML models like Random Forest demonstrated strong performance, their limited interpretability compared to simpler models like Logistic Regression poses challenges for clinical adoption. Future research should focus on validating these findings across diverse datasets to enhance generalizability. Addressing outcome imbalances through advanced techniques without data loss will be critical for better prediction of rare outcomes like hospice discharge or leaving against medical advice. Moreover, future studies could explore fitting a Generalized Additive Model with a Zero-Inflated Negative Binomial (ZINB) distribution to account for overdispersion and zero inflation in LoS data. Expanding this work to include real-time predictive tools integrated into Clinical Decision Support Systems (CDSS) could provide dynamic insights for resource allocation and care optimization.

5. CONCLUSION

This study highlights the potential of machine learning in advancing critical care for hepatitis patients. The models identified key predictors of LoS, discharge outcomes, and discharge locations, with clinical factors such as pre-ICU and ICU procedures and medication counts playing a significant role in determining these outcomes. Additionally, sociodemographic factors, including age and race, were consistently identified as important predictors, highlighting the presence of racial disparities among ICU-admitted patients diagnosed with hepatitis. These findings highlight the importance of addressing such disparities while using predictive analytics to optimize resource allocation, improve patient outcomes, and guide targeted interventions. Despite its strengths, the study also revealed challenges such as variability in

extended LoS and limited data for multi-class predictions, underscoring the need for further advancements in predictive modeling strategies.

6. CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

7. FUNDING STATEMENT

This study did not receive any funding.

8. ETHICS STATEMENT

The Indiana University Human Research Protection Program (HRPP) staff determined the analysis done in this study was not human subject research and did not require further Institutional Review Board (IRB) review before conducting the study. The study was conducted under the data use agreement governing MIMIC-IV, ensuring compliance with ethical standards for research involving human subjects. The collection of patient information and creation of the research resource was reviewed by the Institutional Review Board at the Beth Israel Deaconess Medical Center, who granted a waiver of informed consent and approved the data sharing initiative.

9. DATA AVAILABILITY STATEMENT

The datasets used for this study are available at MIMIC-IV v3.0 only for credentialed users following the PhysioNet Credentialed Health Data Use Agreement (DUA).

References

[1] Abad, Z. S. H., Maslove, D. M., & Lee, J. (2021). Predicting Discharge Destination of Critically Ill Patients Using Machine Learning. IEEE journal of biomedical and health informatics, 25(3), 827–837. https://doi.org/10.1109/JBHI.2020.2995836

[2] Alghatani, K., Ammar, N., Rezgui, A., & Shaban-Nejad, A. (2021). Predicting intensive care unit length of stay and mortality using patient vital signs: Machine learning model development and validation. JMIR Medical Informatics, 9(5), e21347. https://doi.org/10.2196/21347

[3] An, Q., Rahman, S., Zhou, J., & Kang, J. J. (2023). A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges. *Sensors*, *23*(9), 4178. https://doi.org/10.3390/s23094178

[4] Arnab Kumar Ghosh, Geisler, B. P., & Ibrahim, S. A. (2021). Racial/ethnic and socioeconomic variations in hospital length of stay. *Medicine*, *100*(20), e25976–e25976. https://doi.org/10.1097/md.0000000000025976

[5] Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. https://doi.org/10.1002/mpr.329

[6] Brunner, N., & Bruggmann, P. (2021). Trends of the global Hepatitis C disease burden: Strategies to achieve elimination. Journal of Preventive Medicine and Public Health, 54(4). https://doi.org/10.3961/jpmph.21.151

[7] Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations inR. Journal of Statistical Software, 45(3). https://doi.org/10.18637/jss.v045.i03

[8] Centers for Disease Control and Prevention. (2024, April 3). *2022 Viral Hepatitis Surveillance Report | CDC*. https://www.cdc.gov/hepatitis/statistics/2022surveillance/index.htm

[9] Choi, M. H., Kim, D., Choi, E. J., Jung, Y. J., Choi, Y. J., Cho, J. H., & Jeong, S. H. (2022). Mortality prediction of patients in intensive care units using machine learning algorithms based on electronic health records. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-11226-4

[10] Couronné, R., Probst, P., & Boulesteix, A. L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, *19*(1), 270. https://doi.org/10.1186/s12859-018-2264-5

[11] Dubin, J. A., Bains, S. S., Rubén Monárrez, Gilmor, R., Swartz, G. N., Katanbaf, R. M., Mont, M. A., Nace, J., & Delanois, R. E. (2024). The Effect of Insurance Type on Length of Stay Following Total Knee Arthroplasty. *The Journal of Arthroplasty*. https://doi.org/10.1016/j.arth.2024.07.009

[12] Eskandari, M., Alizadeh Bahmani, A. H., Mardani-Fard, H. A., Karimzadeh, I., Omidifar, N., & Peymani, P. (2022). Evaluation of factors that influenced the length of hospital stay using data mining techniques. *BMC Medical Informatics and Decision Making*, *22*(1). https://doi.org/10.1186/s12911-022-02027-w

[13] Fernandez, G. A., & Vatcheva, K. P. (2022). A comparison of statistical methods for modeling count data with an application to hospital length of stay. *BMC Medical Research Methodology*, *22*(1). https://doi.org/10.1186/s12874-022-01685-8

[14] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.

[15] Gwynn, M. E., Poisson, M. O., Waller, J. L., & Newsome, A. S. (2019). Development and validation of a medication regimen complexity scoring tool for critically ill patients. American Journal of Health-System Pharmacy, 76(Supplement_2), S34–S40. https://doi.org/10.1093/ajhp/zxy054

[16] Hastie, T. and Tibshirani, R. (2014). Generalized Additive Models† . In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). https://doi.org/10.1002/9781118445112.stat03141

[17] Hayes, R. M., Carter, P. R., Gollop, N. D., Reynolds, J., Uppal, H., Sarma, J., Chandran, S., & Potluri, R. (2016). 108 The Impact of Marital Status on Mortality and Length of Stay in Patients Admitted with Myocardial Infarction: Abstract 108 Table 1. *Heart*, *102*(Suppl 6), A77.1-A77. https://doi.org/10.1136/heartjnl-2016-309890.108

[18] Hickman, D. (2018, January 3). *Choosing location after discharge wisely*. The Hospitalist. https://www.the-hospitalist.org/hospitalist/article/155424/transitions-care/choosing-location-after-discharge-wisely

[19] Ismail, N., & Jemain, A. (2007). Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models. https://www.casact.org/sites/default/files/database/forum_07wforum_07w109.pdf

[20] Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L. A., & Mark, R. (2024). MIMIC-IV (version 3.0). PhysioNet. https://doi.org/10.13026/hxp0-hg59.

[21] Johnson, A.E.W., Bulgarelli, L., Shen, L. Et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 10, 1 (2023). https://doi.org/10.1038/s41597-022-01899-x

[22] Kwak, C., & Clayton-Matthews, A. (2002). Multinomial logistic regression. *Nursing research*, *51*(6), 404–410. https://doi.org/10.1097/00006199-200211000-00009

[23] Levin, S., Barnes, S., Toerper, M., Debraine, A., DeAngelo, A., Hamrock, E., Hinson, J., Hoyer, E., Dungarani, T., & Howell, E. (2020). Machine-learning-based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length-of-stay. *BMJ Innovations*, *7*(2), 414–421. https://doi.org/10.1136/bmjinnov-2020-000420

[24] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18–22. https://CRAN.R-project.org/doc/Rnews/

[25] Lingsma, H. F., Bottle, A., Middleton, S., Kievit, J., Steyerberg, E. W., & Marang-van de Mheen, P. J. (2018). Evaluation of hospital outcomes: The relation between length-of-stay, readmission, and mortality in a large international administrative database. BMC Health Services Research, 18(1). https://doi.org/10.1186/s12913-018-2916-1

[26] Mickle, C. F., & Deb, D. (2022). Early prediction of patient discharge disposition in acute neurological care using machine learning. BMC Health Services Research, 22(1). https://doi.org/10.1186/s12913-022-08615-w

[27] Moitra, V. K., Guerra, C., Linde-Zwirble, W. T., & Wunsch, H. (2015). Relationship between ICU length of stay and long-term mortality for elderly ICU survivors. *Critical Care Medicine*, *44*(4), 1. https://doi.org/10.1097/ccm.0000000000001480

[28] Ng, M., Patrizia Maria Carrieri, Lindila Awendila, Maria Eugenia Socías, Knight, R., & Ti, L. (2024). Hepatitis C virus infection and hospital-related outcomes: A systematic review. *Canadian Journal of Gastroenterology & Hepatology*, *2024*, 1–23. https://doi.org/10.1155/2024/3325609

[29] Olanipekun, T., Abe, T., Sobukonla, T., Tamizharasu, J., Gamo, L., Kuete, N. T., Bakinde, N., Westney, G., & Snyder, R. H. (2021). Association between race and risk of ICU mortality in mechanically ventilated COVID-19 patients at a safety net hospital. Journal of the National Medical Association. https://doi.org/10.1016/j.jnma.2021.09.003

[30] Oracle Corporation. (2023). MySQL: The world's most popular open-source database (Version 8.0) [Computer software]. Oracle. https://www.mysql.com/

[31] Oracle. (2021). *What is MySQL?* https://www.oracle.com/mysql/what-is-mysql/

[32] Peres, I. T., Hamacher, S., Oliveira, F. L. C., Thomé, A. M. T., & Bozza, F. A. (2020). What factors predict length of stay in the intensive care unit? Systematic review and meta-analysis. *Journal of Critical Care*, *60*, 183–194. https://doi.org/10.1016/j.jcrc.2020.08.003

[33] Petruzziello, A., Marigliano, S., Loquercio, G., Cozzolino, A., & Cacciapuoti, C. (2016). Global epidemiology of hepatitis C virus infection: An up-date of the distribution and circulation of hepatitis C virus genotypes. World Journal of Gastroenterology, 22(34), 7824. https://doi.org/10.3748/wjg.v22.i34.7824

[34] Pradipta, G. A., Wardoyo, R., Musdholifah, A., Sanjaya, I. N. H., & Ismail, M. (2021). SMOTE for handling imbalanced data problem: A review. *Proceedings of the 2021 Sixth International Conference on Informatics and Computing (ICIC)*, 1–8. https://doi.org/10.1109/ICIC54025.2021.9632912

[35] R Core Team. (2023). R: A language and environment for statistical computing (Version 4.4.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

[36] RStudio Team. (2023). RStudio: Integrated development for R (Version 2024.4.2.764) [Computer software]. RStudio, PBC. https://www.rstudio.com/

[37] Smith, B. J. (2024). MachineShop: Machine learning models and tools (Version 3.8.0) [R package]. CRAN. https://cran.r-project.org/package=MachineShop

[38] Stoltzfus, J. C. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine*, *18*(10), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

[39] Tola Getachew Bekele, Melaku, B., Lemlem Beza Demisse, Legese Fekede Abza, & Awol Seid Assen. (2024). Outcomes and factors associated with prolonged stays among patients admitted to adult intensive care unit in a resource-limited setting: A multicenter chart review. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-64911-x

[40] Torgo, L. (2010). Data mining with R: Learning with case studies. Chapman and Hall/CRC. Retrieved from http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR

[41] Venables, W. N., & Ripley, B. D. (2002). Modern applied statistics with S (4th ed.). Springer. https://doi.org/10.1007/978-0-387-21706-2

[42] Verma, V.K., Saxena, K., Banodha, U. (2024). Analysis Effect of K Values Used in K Fold Cross Validation for Enhancing Performance of Machine Learning Model with Decision Tree. In: Garg, D., Rodrigues, J.J.P.C., Gupta, S.K., Cheng, X., Sarao, P., Patel, G.S. (eds) Advanced Computing. IACC 2023. Communications in Computer and Information Science, vol 2053. Springer, Cham. https://doi.org/10.1007/978-3-031-56700-1_30

[43] Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag. https://ggplot2.tidyverse.org

[44] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), Article 1686. https://doi.org/10.21105/joss.01686

[45] Wilimitis, D., & Walsh, C. G. (2023). Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial. *JMIR AI*, *2*(1), e49023. https://doi.org/10.2196/49023

[46] Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

[47] World Health Organization. (2023, October 27). *Hepatitis*. Www.who.int. https://www.who.int/health-topics/hepatitis/viral-hepatitis-in-the-world-map#tab=tab_1

[48] World Health Organization. (2024, April 9). *WHO sounds alarm on viral hepatitis infections claiming 3500 lives each day*. https://www.who.int/news/item/09-04-2024-who-sounds-alarm-on-viral-hepatitis-infections-claiming-3500-lives-each-day

[49] Wrld, A. (2024, September 10). *Random Forest: The Ultimate Guide to Regression and Classification*. Medium. https://medium.com/@adeevmardia/random-forest-the-ultimate-guide-to-regression-and-classification-33506d6cf865

[50] Xu, Z., Zhao, C., Scales, C. D., Henao, R., & Goldstein, B. A. (2022). Predicting in-hospital length of stay: a two-stage modeling approach to account for highly skewed data. BMC Medical Informatics and Decision Making, 22(1). https://doi.org/10.1186/s12911-022-01855-0

[51] Yoshida, K., & Bartel, A. (2022). tableone: Create 'Table 1' to describe baseline characteristics with or without propensity score weights (Version 0.13.2) [R package]. CRAN. https://CRAN.R-project.org/package=tableone

[52] Zeileis, A., Kleiber, C., & Jackman, S. (2015). *Regression Models for Count Data in R.* The Comprehensive Archive R Network. https://cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf

[53] Zhang, Z., Zhao, Y., Canes, A., Steinberg, D., Lyashevska, O., & written on behalf of AME Big-Data Clinical Trial Collaborative Group (2019). Predictive analytics with gradient boosting in clinical medicine. *Annals of translational medicine*, *7*(7), 152. https://doi.org/10.21037/atm.2019.03.29

[54] Zwanenburg, A. (2022, December 16). *Performance metrics*. R-Project.org. https://cran.r-project.org/web/packages/familiar/vignettes/performance_metrics_precompiled.html