Predicting ICU Readmissions Using Electronic Health Records: A Machine Learning Approach

Joseph Lattanzi
Master of Science in Applied Data Science
Case Analysis Capstone
09/01/2025

Problem Statement:

Hospital readmissions represent a critical challenge in healthcare, with readmission rates remaining persistently high despite targeted interventions. Approximately 20% of Medicare beneficiaries experience readmission within 30 days, with average US hospital readmission rates of 14.67% across all conditions (CMS, 2044; Definitive Healthcare, 2024). Hospital readmission costs approximately $26 billion annually in the United States, prompting Medicare to implement penalties of up to 3% of total Medicare payments for hospitals with excessive readmission rates (Commonwealth Fund, 2023).

Research Question:

Can we develop a predictive model using electronic health record data to identify patients at high risk for ICU readmission within 30 days of discharge?

Clinical Significance:

Early identification of high-risk patients could enable targeted interventions, improve patient outcomes, and optimize resource allocation. This project addresses a real-world healthcare problem with direct applications for hospital quality improvement initiatives.

Dataset Description:

Primary Dataset: MIMIC-IV (Medical Information Mart for Intensive Care IV)

- Source: MIT Laboratory for Computational Physiology, PhysioNet
- Size: 73,181 patients, 523,740 hospital admissions
- Time Period: 2008-2019 from Beth Israel Deaconess Medical Center
- Data Type: De-identified electronic health records

Key Tables and Variables

- Admissions: 523,740 records with admission/discharge dates, demographics
- Patients: Age, gender, mortality indicators
- Diagnoses_ICD: ICD-9/10 diagnosis codes for comorbidity assessment
- Prescriptions: Medication data for polypharmacy analysis
- Labevents: >100 million laboratory measurements for physiological markers
- Procedures_ICD: Medical procedures performed during stay

Target Variable: Binary readmission indicator (readmitted to ICU within 30 days: yes/no)

Predictor Variables:

- Demographics (age, gender)
- Length of stay
- Discharge laboratory values (albumin, creatinine, hemoglobin, etc.)
- Comorbidity burden (Charlson Comorbidity Index)
- Medication count at discharge
- Procedures performed
- Admission diagnosis category

Methodology and Analysis Plan:

Phase 1: Data Exploration and Preprocessing

- Explore data analysis of patient demographics and admission patterns
- Readmission rate calculation and temporal analysis
- Missing data assessment and imputation strategy
- Feature Engineering:
    - Laboratory value trends (last 24 hours before discharge)
    - Comorbidity scoring
    - Length of stay categorization
    - Medication complexity metrics

Phase 2: Model Development

- Baseline Model: Logistic regression for interpretability
- Advanced Models: Random Forest, Gradient Boosting (XGBoost)
- Feature Selection: Clinical knowledge-guided and statistical methods
- Validation Strategy: Time-aware cross-validation to prevent data leakage

Phase 3: Model Evaluation and Interpretation

- Performance Metrics:
    - Area Under the ROC Curve (AUC) - primary metric

    - Sensitivity, specificity, positive/negative predictive values
    - Calibration assessment for clinical utility
- Feature Importance Analysis: Identify key predictors for clinical insights
- Subgroup Analysis: Performance across different patient populations

Phase 4: Clinical Translation and Presentation

- Risk stratification framework development
- Clinical interpretation and recommendations
- Comparison to existing readmission prediction tools
- Final presentation and technical documentation

Timeline:

Week	Activity	Deliverable
1-2	Data acquisition, EDD, data wrangling	Data summary report
3-4	Feature engineering, preprocessing, cleaning	Clean modeling dataset
5-6	Model development and tuning	Trained models
7-8	Evaluation, interpretation, and presentation prep	Final presentation

Expected Outcomes and Success Metrics:

Technical Goals:

- Achieve AUC > 0.75 (clinically meaningful prediction performance)
- Identify 5-7 key predictive features for readmission risk
- Develop interpretable risk scoring system

Clinical Impact:

- Provide actionable insights for discharge planning
- Identify modifiable risk factors for intervention
- Demonstrate potential for integration into hospital workflows

Technical Requirements:

Software: R with key packages:

- Data manipulation: dplyr, tidyr, data.table
- Visualization: ggplot2, plotly for interactive plots
- Machine Learning: caret, randomForest, gbm, glmnet
- Statistical Analysis: pROC, rms, VIM, Hmisc

- Reporting: R Markdown, knitr

Infrastructure: RStudio environment with standard statistical computing capabilities
Ethical Considerations: All data is de-identified; following established research protocols

References:

1. Centers for Medicare & Medicaid Services. (2024). Hospital Readmissions Reduction Program: Fiscal Year 2024 Results. CMS.gov.
2. Definitive Healthcare. (2024). Hospital Readmission Rates by State and Hospital. Healthcare Analytics Report.
3. Commonwealth Fund. (2023). Reducing Hospital Readmissions: Current Strategies and Future Directions. Health Policy Brief.
4. Johnson, A. E., Bulgarelli, L., Shen, L., et al. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10, 1.