TITLE: Optimizing Healthcare Efficiency through Casemix Driven Machine Learning: A Predictive Model for High-risk Readmissions

Introduction

Unplanned all cause readmissions not only strain hospital capacity, but also indicate gaps in care coordination and patient management. Leveraging casemix data and machine learning (ML) capabilities offers a promising approach to reduce preventable hospital readmissions, improve healthcare efficiency and improve patient outcomes. This study develops and validates an ML model for the early identification and prediction of patients at risk of readmission and provides methodology to identify hospitals with elevated readmissions rates.

Methods

We used the National Platform for Health and Insurance Exchange Services (NPHIES) claims data as the primary source to identify predictive factors associated with patient and hospital readmission risk. The dataset included approximately 17 million inpatient claims for more than 0.5 M unique patients from 228 hospitals, covering the period between May 2023 and May 2024, excluding potentially planned readmissions. Data preprocessing involved rigorous cleaning procedures to address inconsistencies, duplicates and missing values, with an AI-assisted quality assessment framework employed to rectify inaccuracies in patient demographics and invalid International Classification of Diseases 10th Revision Australian Modification (ICD-10-AM 10th Edition) codes. Feature engineering was performed to extract meaningful historical trends and enhance model performance. Multiple predictive models were developed and assessed, with gradient boosted trees emerging as the optimal model based on key performance metrics. Additionally, based on the expected/observed readmissions ratio, we developed a provider scoring methodology to assess hospital performance.

Results

The model's predictive performance was assessed using multiple evaluation metrics, including overall accuracy, recall, precision, F1-score, and area under the curve of the receiver operating characteristic (AUC-ROC). The best-performing model achieved scores on unseen data of 82.37, 61.18, 12.32, 20.51, and 72.18, respectively. Feature importance analysis, using Shapley Additive Explanations (SHAP), revealed that the top 3 most significant predictors are ICD-10-AM codes, hospital license, and the number of services and procedures provided during the inpatient admission. SHAP analysis also highlighted the interaction effects between length of stay, and number of services provided with readmission risk, reinforcing the need for patient-specific risk assessment. To evaluate provider performance, a risk-adjusted provider scoring methodology was developed. This approach incorporated patient-level risk factors, including age, gender, principal diagnosis, length of stay, presence of secondary diagnoses, to adjust the expected-to-observed readmissions ratio. Hospitals were ranked based on their risk-adjusted readmission performance, allowing for more equitable benchmarking. A two-sided Poisson test was conducted to assess the statistical significance of deviations in provider performance, confirming meaningful variations in hospital-level readmission rates.

Discussion/Conclusions

This study highlights the role of casemix-driven ML models in identifying high risk factors, such as diagnoses codes and hospital characteristics, in predicting 30-days hospital readmissions. Utilizing flexible ML frameworks, such as gradient boosted trees, overcomes issues such as data imbalance, while having the possibility of having real time data

prediction. Further work could explore model performance improvement by integration of comorbidity risks, chronic disease flags, and prior healthcare utilization patterns. Finally, interventions like better discharge planning, care coordination, chronic disease management, and optimizing resources could reduce readmission rates.