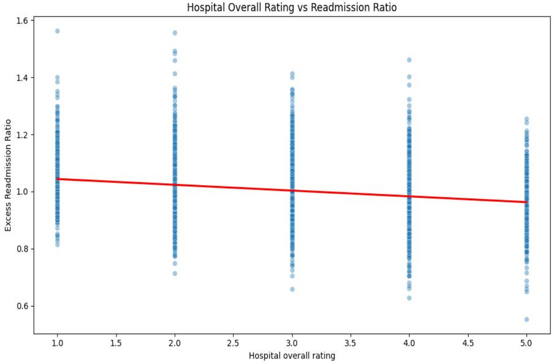
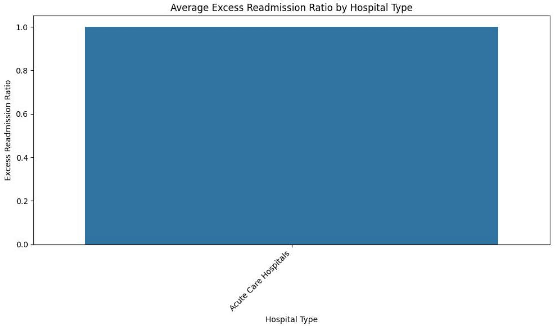
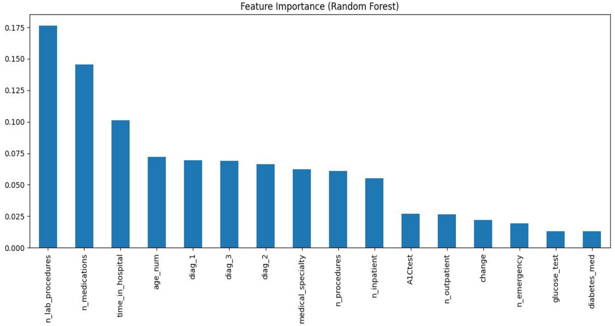
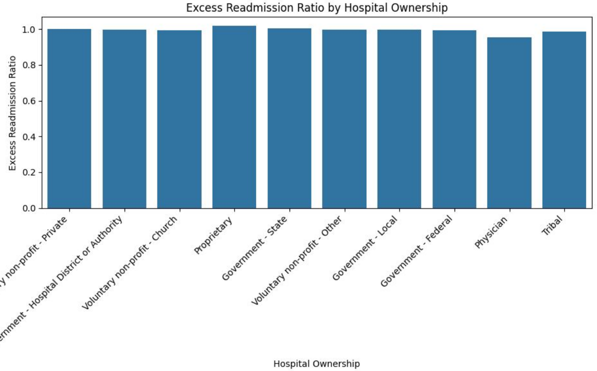


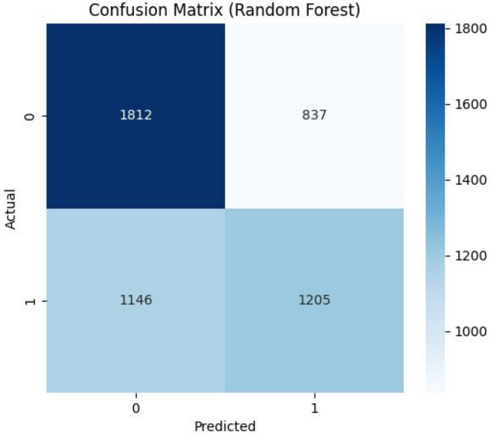


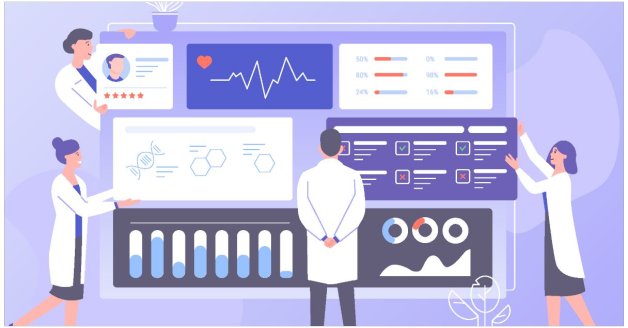


Predicting Hospital Readmissions & Analyzing U.S. Hospital Quality Using CMS Data

DSCI-510 Final Project (Fall 2025)
Warren Lim







Introduction

This project investigates hospital readmission patterns in the United States by combining machine learning with national healthcare quality data. The first component uses a patient-level clinical dataset to predict the likelihood of readmission based on demographic, diagnostic, and treatment-related variables. The second component analyzes national hospital performance and readmission outcomes using publicly available CMS API datasets. Together, these analyses provide both a predictive and systemic understanding of readmissions, highlighting how patient complexity, hospital ownership, and quality ratings relate to readmission trends.

Project Overview

- Explore factors tied to hospital readmissions
- Predict readmission using a Kaggle clinical dataset
- Analyze national hospital quality using CMS HRRP APIs
- Identify trends in ratings, ownership, and outcomes

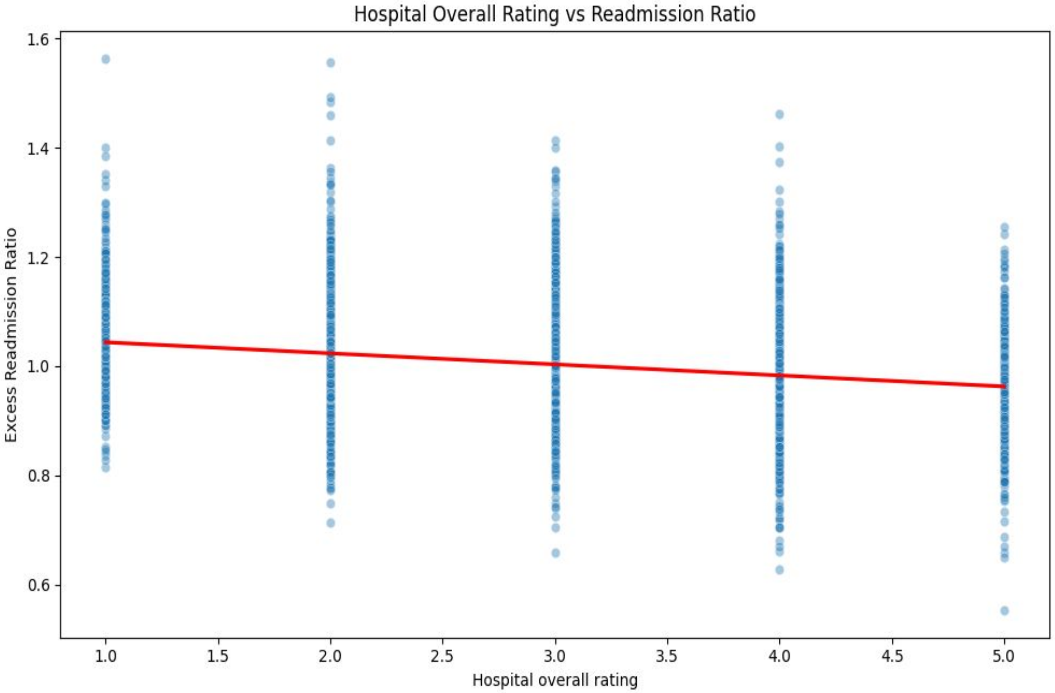
Data Sources

Dataset	Description	Source	Type	Rows
CMS HRRP (Readmissions)	Hospital-level readmission and quality data	https://data.cms.gov/provider-data/dataset/9n3s-kdb3	API	18,510
CMS Hospital General Info	Includes essential data: location, type, ownership, and overall rating for hospitals.	https://data.cms.gov/provider-data/dataset/xubh-q36u	API	5,381
Kaggle Diabetes Readmissions	Patient-level data for predicting 30-day hospital readmission rates based on clinical and demographic features	https://www.kaggle.com/datasets/dubradave/hospital-readmissions	CSV	25,000

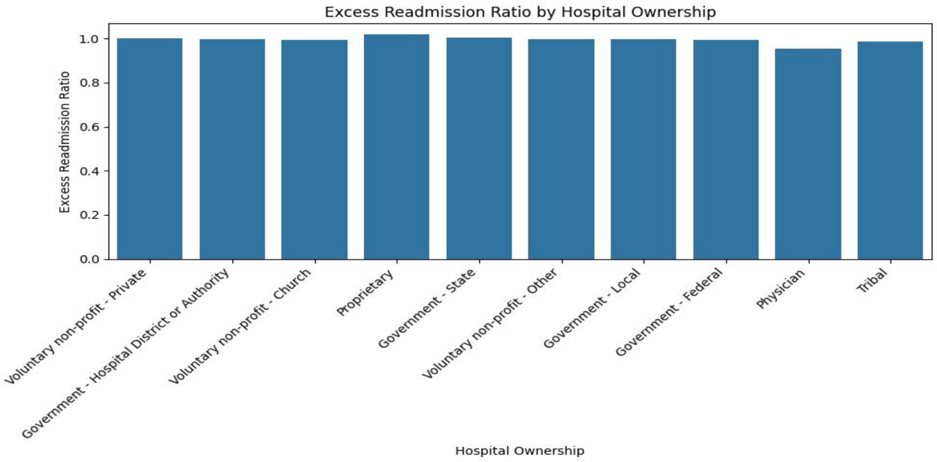
Hospital Rating vs Readmission Ratio

- Weak negative correlation
- Higher rated hospitals → slightly fewer readmissions
- Relationship not very strong

- Here we compare hospital overall ratings to excess readmission ratios. The trend line shows a weak negative correlation—higher rated hospitals generally have lower readmission ratios. However, the variation is large, suggesting readmission outcomes depend on more than the star rating alone.

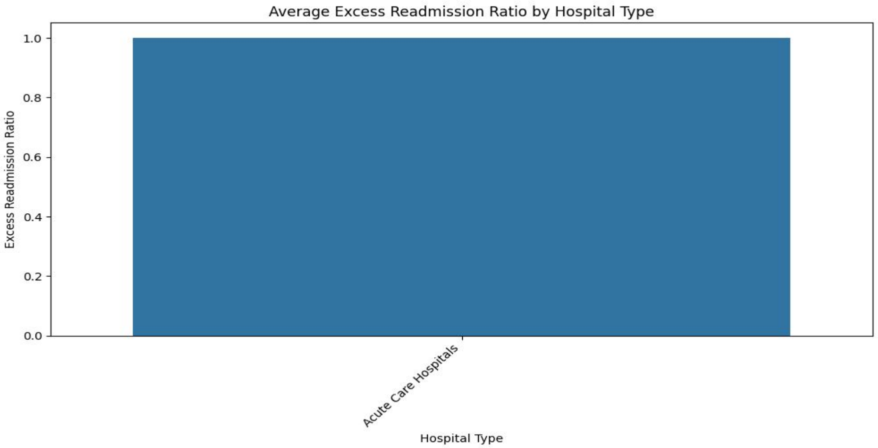


Readmission by Hospital Ownership & Type (CMS)



- For-profit (Proprietary) hospitals → highest readmission
- Non-profit hospitals perform slightly better
- Government hospitals vary but stay close to average

- This plot shows interesting system-level differences. For-profit hospitals have higher readmission rates compared to non-profits and government-owned hospitals. This aligns with known patterns in healthcare quality literature. Still, the differences are moderate, not dramatic.



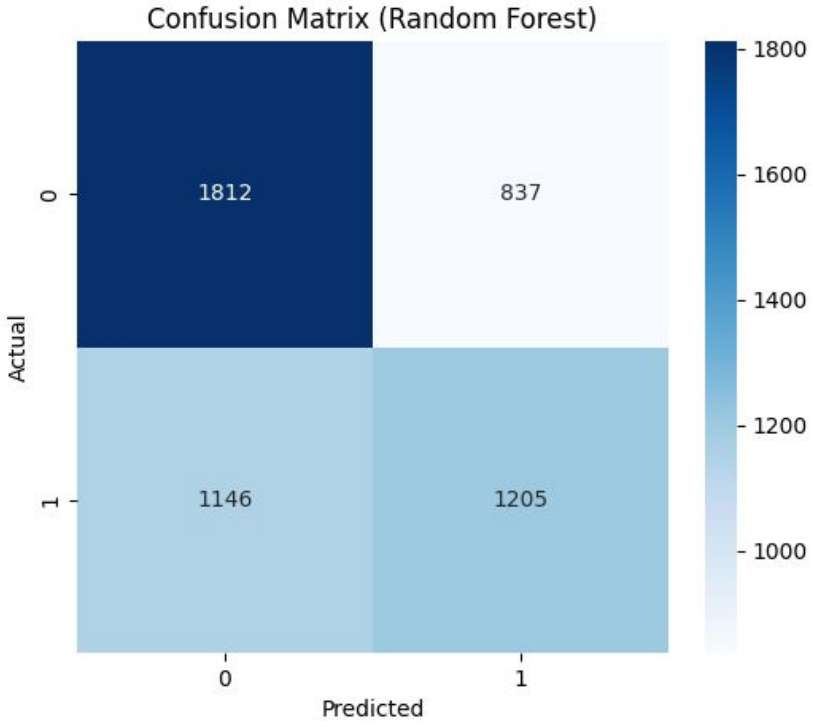
- Almost all hospitals are Acute Care Hospitals
- CMS HRRP primarily evaluates acute-care facilities
- Limited variation by type

- CMS HRRP focuses nearly entirely on acute-care hospitals, which is why this plot shows a single category. This means hospital-type variation is limited in this dataset.

Readmission Prediction (Kaggle Dataset)

- Accuracy of roughly 60% (0.6034)
- Good at predicting “no readmission”
- Struggles with identifying true readmissions

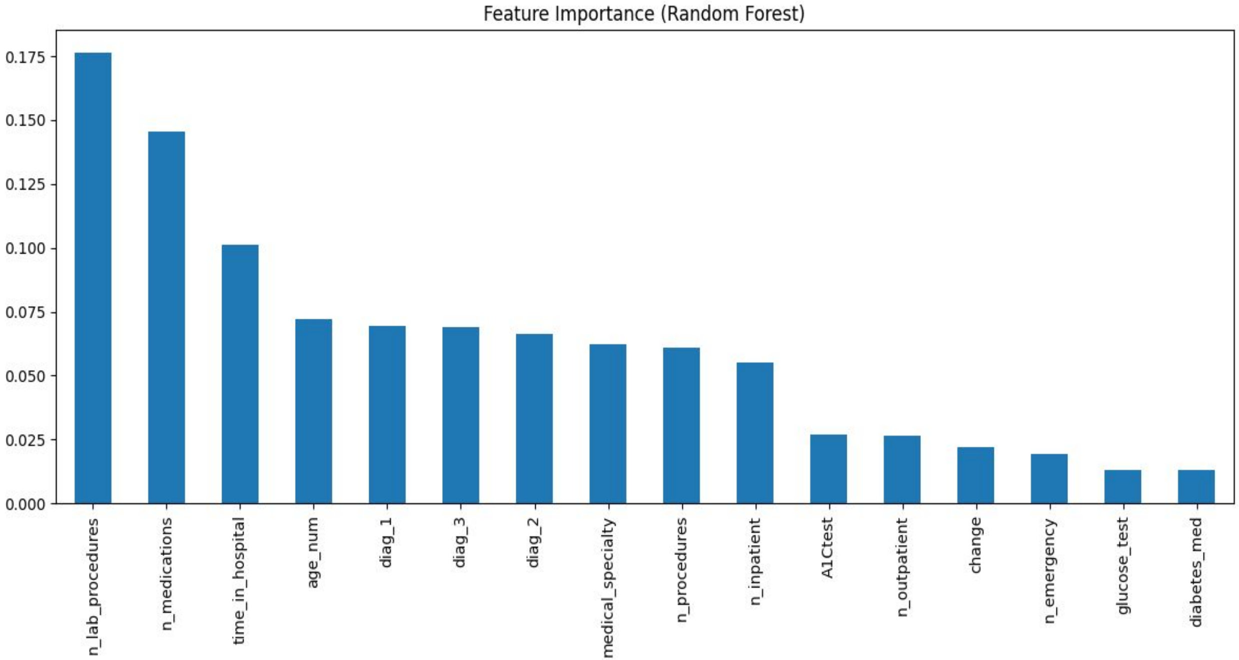
- Here we trained logistic regression and random forest models, achieving around 60% accuracy. The confusion matrix shows the model correctly predicts non-readmissions more often than actual readmissions. This indicates the dataset is somewhat imbalanced and readmissions remain hard to predict.



Top Predictors of Readmission

- Lab procedures and medication count
- Length of stay
- Age and clinical diagnoses
- Complexity drives readmission risk

- Feature importance from the random forest reveals that readmissions are driven mainly by clinical complexity — the number of lab procedures, medication count, and length of stay are the strongest predictors. These correlate with patients being sicker or needing more care.

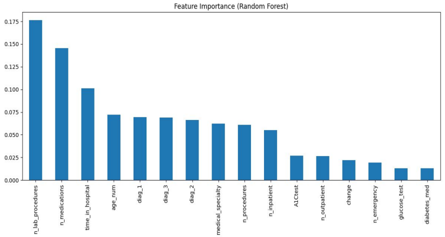


Challenges

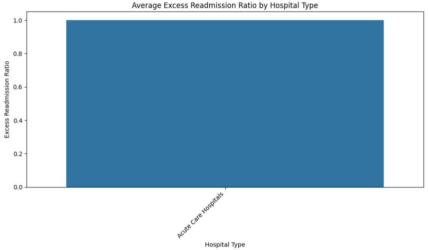
- CMS datasets contain “Not Available” and string-based numeric fields
- Needed data cleaning + numeric conversion
- Dataset merge required key alignment
- Dealing with imbalanced outcomes in ML

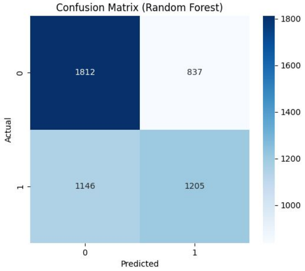
- There were several challenges. CMS datasets use inconsistent formats and often store numeric fields as strings, which required careful cleaning. Merging datasets needed standardization of the Facility ID field. For the Kaggle dataset, balancing readmissions and tuning models was also challenging.

Conclusion

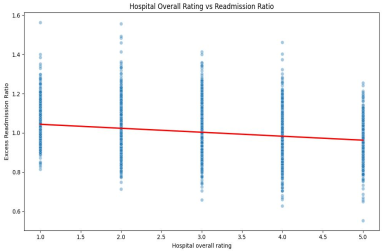


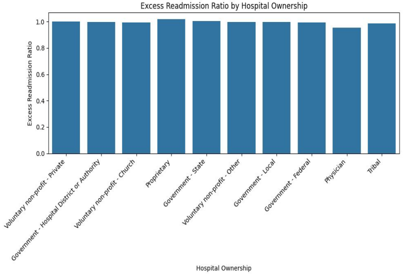
- Readmission outcomes vary more by ownership than rating
- Machine learning models show modest predictive capability
- Clinical complexity is the strongest driver of readmissions
- CMS APIs provide powerful, real-world healthcare insights





- In conclusion, hospital readmission patterns are influenced by systemic factors like ownership, but hospital rating alone is only a weak predictor. Machine learning models showed moderate success, with clinical complexity being the strongest driver of readmissions. Using real healthcare APIs gave a much deeper understanding of nationwide hospital performance.


Hospital Overall Rating vs Readmission Ratio
Excess Readmission Ratio
Hospital overall rating


Excess Readmission Ratio by Hospital Ownership
Excess Readmission Ratio
Voluntary non-profit - Private
Government - Hospital District or Authority
Voluntary non-profit - Church
Proprietary
Government - State
Voluntary non-profit - Other
Government - Local
Government - Federal
Physician
Tribal
Hospital Ownership