PREDICTING AND REDUCING HOSPITAL READMISSIONS IN DIABETIC PATIENTS: A DATA-DRIVEN APPROACH

QMH Fest Data Analysis Competition 2025

Organized by

Team Name : House of Loosers

Team Members :

Name : **Rishima Anzum Risha**
Roll Number : 08
Email : rishima.anzum.risha@gmail.com
Phone : 01330218832

Name : **Himel Sarker Rivu**
Roll Number : 49
Email : hsrivu332@gmail.com
Phone : 01571204165

Name : **Jotirmoy Kumar Roy**
Roll Number : 36
Email : jotirnoyroyy@gmail.com
Phone : 01787917407

Submitted to : **QMH Statistics Club**
Institution : **University of Dhaka**
Submission Date : **October 25, 2025**

This research analyzes hospital readmission patterns among diabetic patients to identify key predictive factors and develop targeted intervention strategies for reducing preventable 30-day readmissions.

Introduction

Unplanned hospital readmissions represent a significant challenge within the modern healthcare landscape. These events are a major concern as they are frequently viewed as indicators of suboptimal care quality. This problem calls for a data-driven approach to reduce preventable readmissions.

The primary objective of this analysis is to conduct a statistical investigation to identify the key factors that are significant drivers of 30 day hospital readmission and to leverage these insights to develop and validate a robust predictive model.

To achieve these objectives various statistical tools, including Exploratory Data Analysis, Classical Hypothesis Testing and Advanced Machine Learning algorithms were applied to find out significant patterns of 30 day readmissions and build a reliable predictive framework.

Research Question & Hypothesis

Our study is guided by the following central research question:

Which demographic, clinical, and healthcare utilization factors most strongly predict 30 day hospital readmissions among patients with diabetes and how can predictive modeling be used to prevent avoidable readmissions?

This question is broad, encompassing various potential variables. To move from general investigation to specific questions we developed five distinct, testable hypotheses. These hypotheses were formulated from initial patterns observed.

Hypotheses:

H1: Patients with both high **medication burden** (>15 medications) and long **hospital stays** (>4 days) experience significantly higher 30 day readmission rates compared to patients with only one or neither factor.

H2: Patients with both frequent prior **inpatient admissions** (>1) and **high emergency department utilization** (>1) have substantially higher 30-day readmission rates than those with only one or neither type of utilization.

H3: Patients who experienced a medication change during hospitalization and had an abnormal A1C result (>7) experience higher 30-day readmission than those with only one or neither factor.

H4: Patients discharged against medical advice (AMA) have higher odds of 30-day readmission than patients discharged home.

H5: Patients with both very frequent **Emergency visits** (≥3) and high comorbidity burden (**diagnosis_total** ≥ 75th percentile) experience higher 30-day readmission compared to those with only one or neither factor.

Summary of Methods

Data Preparation

The initial dataset was cleaned to prepare it for analysis.

- **Engineering target variable:** At first we relabeled the columns by variable name for our convenience. Since our objective is to find factors of 30 day readmission, we modified our target variable V50: **readmission_status**, where **1** represents a readmission within 30 days (<30) and **0** represents no readmission or readmission after 30 days ("NO" or ">30")

- **Handling missing values:** We dropped V6: **body_weight** column since it contains **97%** missing values. We also eliminated V11: **insurance_code** because it contains **52%** missing value and it is weakly associated with 30 day readmission. Another variable V12: **Provider_speciality** was also ignored because it contains **53%** missing values and contains 84 unique values which only adds noise to the dataset.
  We replaced the missing values of the remaining numeric variables with **median** and categorical variables with **"Unknown"**

Data Preparation for Hypothesis Test

We created binary variables associated with our hypotheses that could be formally tested in our statistical models. The following variables were constructed:

- **For H1 (Medication-Intensive Care):** We created **long_los4** to separate patients who stayed in the hospital for more than **4** days. A second flag, **high_meds15**, was made for patients who received more than **15** medications.

- **For H2 (Healthcare Utilization):** We created **high_inpt1** for patients with more than one prior inpatient visit. We also made **high_ed1** for those with more than one prior emergency visit.
- **For H4 (Medication Instability):** We made **a1c_abn** for patients with abnormal A1C test results (">7" or ">8"). Another flag, **medchg**, was created for patients whose diabetes medications were changed during their stay.

- **For H6 (AMA Discharge):** To compare specific discharge paths, we created **ama_flag** for patients who left "**Against Medical Advice**" (e.g., code 7). We also created a **home_flag** for those discharged to home (e.g., code 1).
- **For H7 (High-Risk Synergy):** We made high_ed3 for patients with 3 or more emergency visits. A second flag, **q3_comorb**, was created for patients with a high number of diagnoses (at or above the 75th percentile).

Finally, we converted the **age_band** categories (like [70-80)) into a single, scaled numeric variable (**age_z**) for the models to use.

Exploratory Data Analysis (EDA)

A comprehensive EDA was conducted to understand the data's structure and identify initial relationships between predictors and the **readmitted_30** target.

Univariate & Demographic Analysis

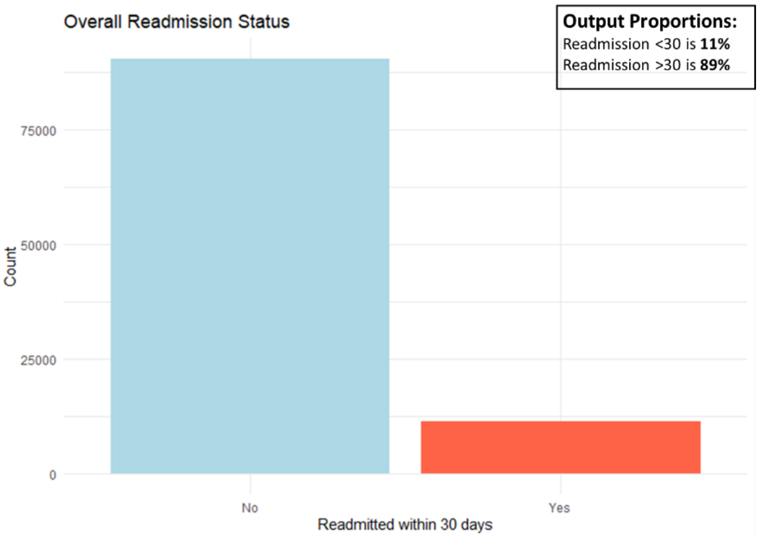
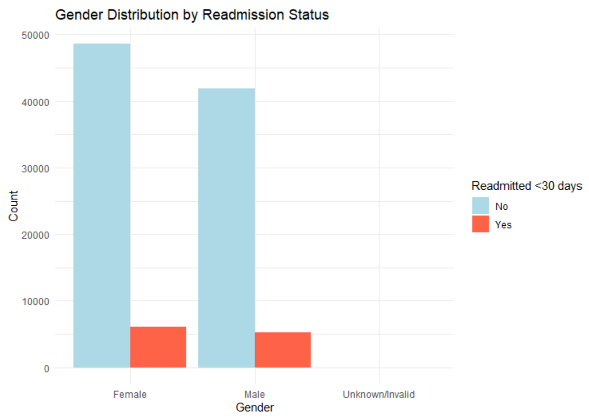


Fig-1: Overall Readmission Status

The Overall Readmission Status (Fig-1) plot immediately revealed a critical feature of the dataset: **severe class imbalance**. The "No" (0) class, representing patients not readmitted within 30 days, makes up the vast majority of encounters (89%). The "Yes" (1) class is a small minority (11%).



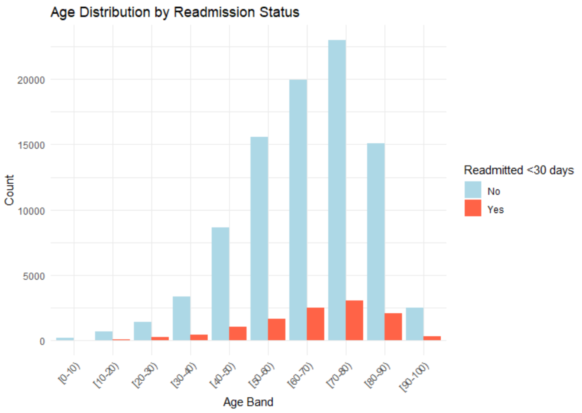


Fig-2: Gender Distribution by Readmission Status **Fig-3:** Age Distribution by Readmission Status

Analysis of patient demographics showed that while the Gender Distribution (Fig-2) was relatively balanced between males and females. The Age Distribution (Fig-3) is positively skewed, 30 day readmissions are heavily concentrated in older age bands, particularly [70-80) and [80-90) This indicates that age could be a significant factor.

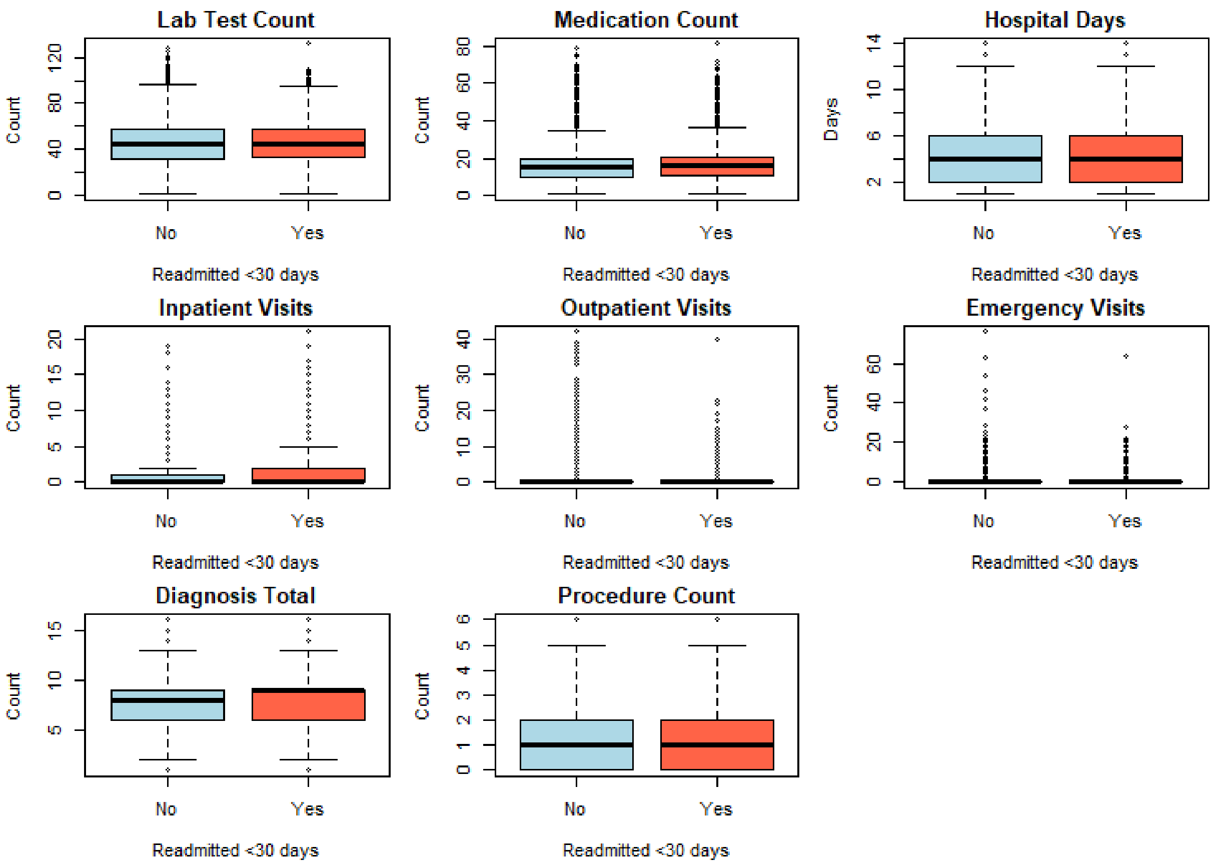


Fig-4: Numerical Variables Distribution by Readmission Status

Following demographics, the distributions of key numeric and categorical variables were compared between the readmitted and non-readmitted groups. The boxplots in Fig-4 provide a visual comparison for numeric variables:

- **Prior Healthcare Utilization:** The most dramatic differences are seen in prior utilization. The median and interquartile range (IQR) for **inpatient_visits** and **emergency_visits** are substantially higher for the "Yes" (readmitted) group.

- **Current Visit Complexity:** Metrics related such as **hospital_days** (length of stay), **medication_count**, and **lab_test_count**, are all visibly higher for patients who were readmitted.

- **Diagnosis Total:** The **diagnosis_total** is also clearly higher for the readmitted group, strongly suggesting that a higher comorbidity burden is associated with readmission risk.

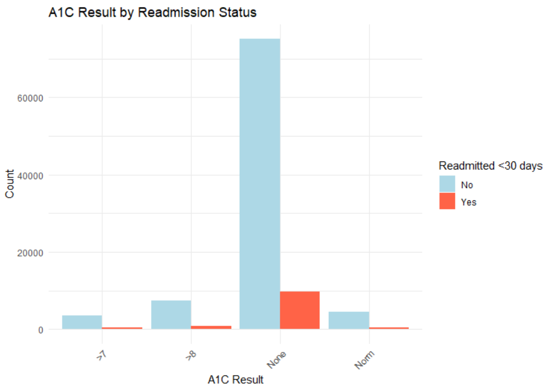


Fig-5: A1C Results by Readmission Status

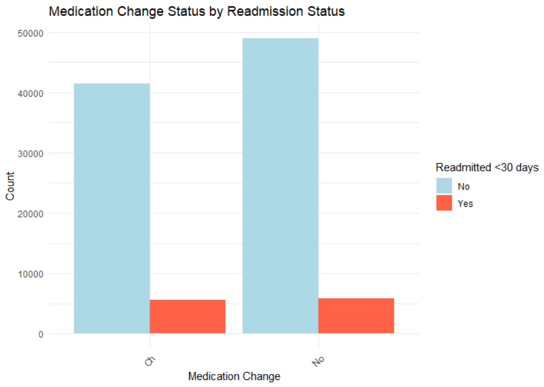


Fig-6: Medication Change Status by Readmission Status

A similar investigation of key categorical variables (Fig-5, Fig-6) reinforced these findings:

- **Clinical Instability:** The A1C Result plot (Fig-5) shows surprising results. It indicates that patients who did not test A1C level have higher readmission rate. Also see that, patients whose medication was changed have higher readmission rate.

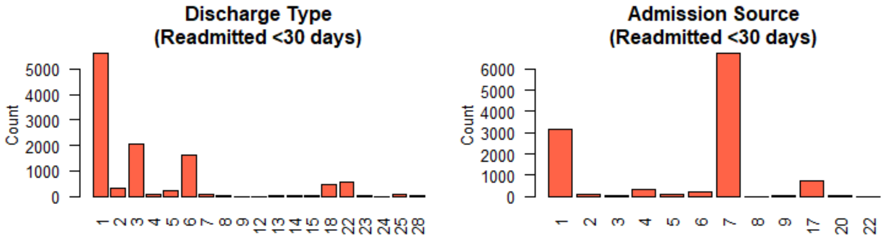


Fig-7&8: Discharge Type & Admission Source by Readmitted patients

- **Discharge Type:** The discharge type graph (EDA7) shows that patients were discharged to home (**discharge_type**=1) have higher readmission rate.

Bivariate Analysis:

To check statistical significance, we applied T-test for numeric variables and Chi-Square test for categorical variables against the target variable. The results are shown below

Numerical Variables	P-Value
lab_test_count	3.89e-11
medication_count	1.32e-34
hospital_days	8.55e-44
inpatient_visits	2.76e-261
outpatient_visits	4.12e-09
emergency_visits	5.39e-42
diagnosis_total	2.07e-64
procedure_count	5.26e-05

Table-1: T-test Table

Catagorical Variables	P-Value
med_change_status	5.21e-10
A1C_result	8.29e-08
age_band	6.60e-21
adm_source_code	8.34e-06
discharge_type	3.53e-319
adm_type_code	3.84e-04

Table-2: Chi-square Table

Numeric Variables: Although all key numeric predictors were found to be profoundly associated with readmission (p-values < 0.05), we can see that **impatient_visits, diagnosis_total** & **hospital_days** are the most important numeric predictors.

Categorical Variables: The categorical variables were similarly significant. The association for **discharge_type** (p=3.53e−319) was very strong, confirming it as a critical factor. Other key variables, **med_change_status** (p=5.21e−10), and **A1C_result** (p=8.29e−08), were all also highly significant.

After profiling individual variables, the analysis progressed to examine the relationships between them. A correlation heatmap (Fig-5) was generated to visualize the linear relationships between the numeric predictors and the **readmitted_30** target.

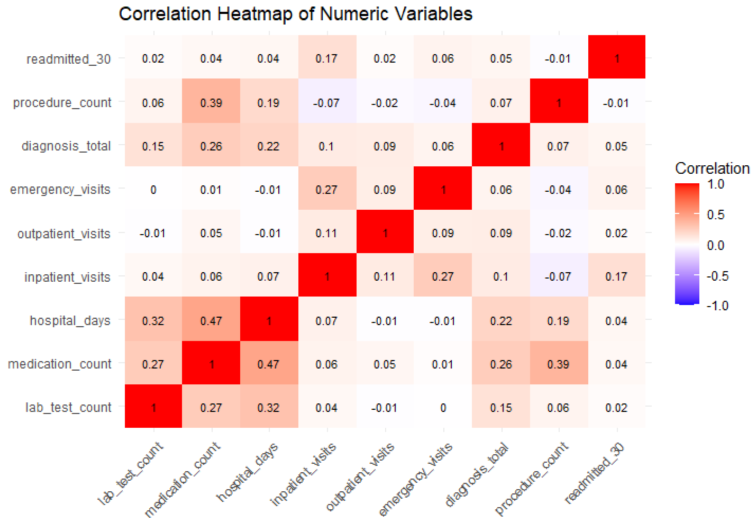


Fig-9: Correlation Heatmap

The heatmap confirms that **inpatient_visits** has the strongest positive correlation with **readmitted_30**, followed by **emergency_visits**, **diagnosis_total**, **hospital_days**, and **medication_count**. This aligns perfectly with the boxplot observations. Furthermore, the heatmap reveals moderate correlations between predictors, such as the relationship between **hospital_days** and **medication_count** (r=0.47), which provides initial support for our H1 hypothesis.

Model Design for prediction

To predict our outcome, we built and tested four different models.

1. Splitting The Data

First, we randomly mixed up all our data one time. Then, we split it into three separate groups to build and test our models fairly:

- **Training Set (70%):** The largest group, used to teach the models.
- **Validation Set (20%):** Used to fine-tune the models and pick the best settings.
- **Test Set (10%):** The final, unseen group used to check how well the models performed.

2. The Models for Testing

We designed four different types of models:

1. **Logistic Regression:** A standard statistical model. We gave extra importance (weight) to the less common group to help the model learn better.
2. **Elastic Net:** A model that is good at selecting the most important factors. We used a standard 5-fold testing process (cross-validation) to find its best settings.
3. **Random Forest**: A model that works by building 600 separate "decision trees" and combining their answers to make a final prediction.
4. **XGBoost:** A powerful and popular model. We carefully adjusted its settings (like learning rate) and also balanced the groups to improve its accuracy.

3. Tuning the Models

For each model, we needed to find the best cut-off point or "threshold" to decide when to predict "yes" or "no." We used the validation set to find the specific threshold for each model that gave the best possible F1-score which is a measure that balances accuracy and completeness.

We then locked in that best threshold and used it when we ran the model on the final test set.

4. How We Measured Performance

To judge which model was best, we reported the following five measurements for each one on the test set:

- **Accuracy**
- **Precision**
- **Recall**

- **F1-Score**
- **AUC (Area Under the Curve)**

Hypothesis Testing by Models

To formally test our five hypothesis we built a predictive tool which is designed to balance statistical interpretability with predictive validity.

1. **Primary Check using GLM :** We used results from our weighted Generalized Linear Model (GLM) to evaluate each hypothesis against a specific set of criteria. A hypothesis was considered **"Supported"** only if it met a strict statistical threshold (**$p < 0.005$**) and demonstrated a positive risk effect (**OR > 1**). A hypothesis was considered having a **"Negative Association"** if it was statistically significant (**$p < 0.05$**) but had an **OR < 1**, indicating a real effect in the opposite direction of our hypothesis. Any hypothesis that did not meet either of these first two conditions was deemed **"Not Supported"**.
2. **Secondary Check using ENet :** A secondary check was performed using an Elastic Net (ENet) model. For a hypothesis to pass this check, the ENet model was required to provide **an importance score above zero**. This confirms that the factor is not just statistically significant but also contributes positively to readmission risk in a predictive model.

A hypothesis was only considered **"Overall Supported"** if it successfully passed both of these independent tests.

Key Results and Model performance

Key Factors

The graphs below show the key factors that drive 30 day readmission among patients with diabetes

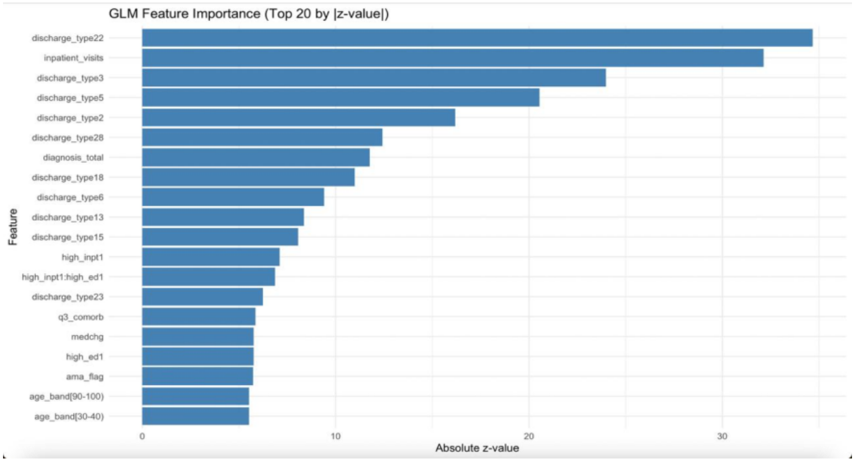


Fig-10: A1C Results by Readmission Status

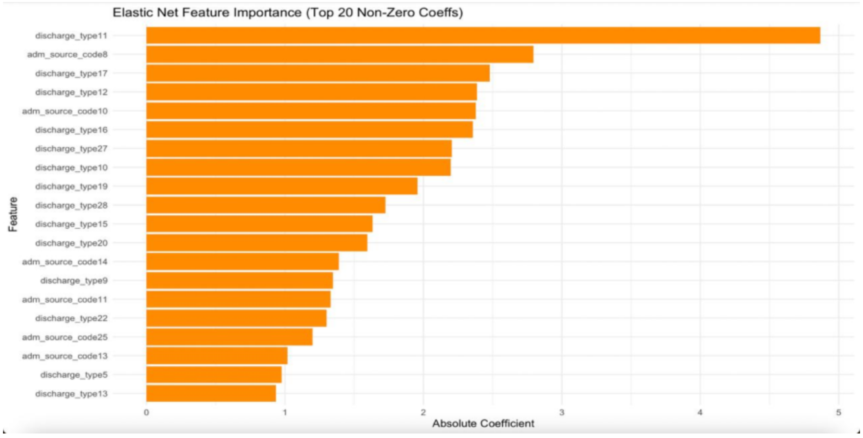


Fig-11: A1C Results by Readmission Status

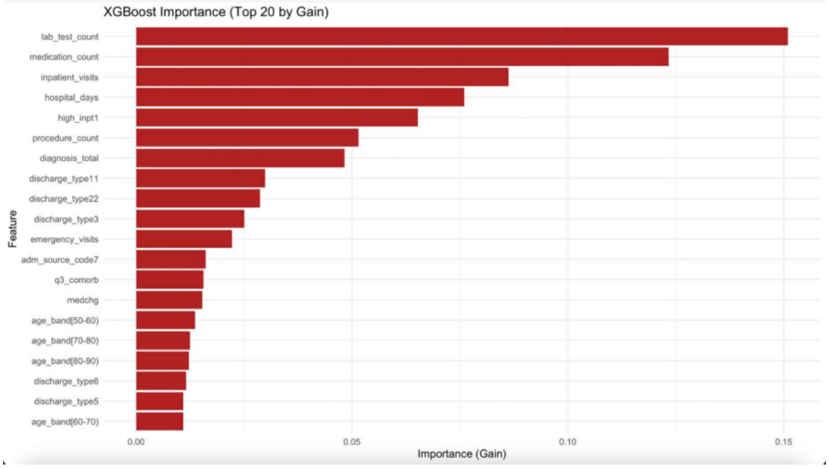


Fig-12: A1C Results by Readmission Status

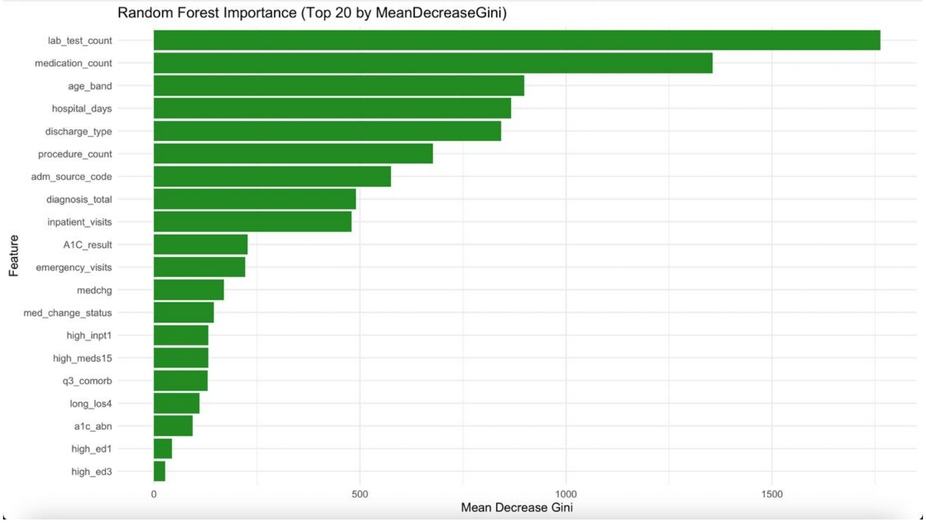


Fig-13: A1C Results by Readmission Status

The feature importance plots derived from all four predictive models highlight several overlapping patterns. When multiple models identify the same variables as influential, these can be considered more trustworthy indicators of 30-day hospital readmission risk.

Discharge Type

Patients discharged home without adequate support or those leaving Against Medical Advice (AMA) are more likely to return within 30 days compared to those discharged to structured care environments such as skilled nursing or rehabilitation facilities.

Prior Utilization (Inpatient and Emergency Visits)

Patients with a history of frequent hospitalizations or emergency visits are more likely to experience another readmission.

Hospital Days or Length of Stay

Longer hospital stays often indicate more complex or severe medical conditions, which in turn increase the likelihood of readmission.

Diagnosis Total

Patients with multiple comorbid conditions face greater challenges in recovery, medication management, and follow-up care, all of which elevate the risk of readmission.

Testing and Treatment Volume (Lab and Medication Counts)

Patients who undergo a higher volume of lab tests and take more medications tend to have more complicated clinical profiles.

To summarize, ten key factors emerged repeatedly across models as the most reliable predictors of 30-day readmission:

Discharge Type, Inpatient Visits, Hospital Days, Diagnosis Total, Lab Test Count, Medication Count, Admission Source Code, Emergency Visits, A1C Result, Medication Change

Predictive Model performance

After constructing four predictive models, each was evaluated on the unseen 10% test set to provide an unbiased assessment of its performance. The primary goal was to determine which model would be most effective for identifying patients at risk of a 30-day readmission.

The performance of the models is summarized in the table below:

Model	Accuracy	Precision	Recall	F1	AUC
GLM	0.713	0.194	0.491	0.279	0.675
Enet	0.713	0.194	0.486	0.278	0.675
RF	0.829	0.235	0.227	0.231	0.640
XGB	0.776	0.213	0.366	0.269	0.651

Table-2: Model Performance Table

The table above summarizes the final performance of the four predictive models on the unseen test set. It highlights a trade-off: the Random Forest (RF) model achieved the highest Accuracy (0.829) which is misleading due to our class imbalanced data. Conversely, the Generalized Linear Model (GLM) and Elastic Net (ENet) models achieved the highest scores for AUC (0.675)

Hypothesis Test Result

Hypothesis	GLM_OR	GLM_P	GLM_Result	ENet_Support	Overall
H1	1.06	2.18e-02	Supported	Supported	**Supported**
H2	0.66	7.29e-12	Negatively associated	Not supported	Negatively associated
H3	0.98	6.33e-01	Not supported	Not supported	Not supported
H4	1.46	1.51e-10	Supported	Supported	**Supported**
H5	0.75	1.10e-04	Negatively associated	Not supported	Negatively associated

Table-3: Hypothesis Test Result

Here,

H1: High meds (>15) × Long stay (>4d) interaction

H2: High inpatient (>1) × High ED (>1) interaction

H3: Abnormal A1C × Med change interaction

H4: AMA discharge vs Home (subset)

H5: High ED (≥3) × High Comorbidity (Q3) interaction

The table above presents the formal statistical findings for our five research hypotheses as per our dual-validation methodology.

The results show that two of our five hypotheses, **H1 (Medication-Intensive Care)** and **H4 (AMA Discharge Risk)**, were fully supported by the data. H2 and H5 have significant p-value but the Odds Ratio indicates that the factors are negatively associated. H3 is not supported at all.

Interpretation of Findings

H1: High meds (>15) and Long stay (>4d) (GLM_OR 1.06)

When a patient presents both a very high medication count and a hospital stay longer than four days, the odds of 30-day readmission are about **6% higher** than what would be expected from the separate effects of medication burden and length of stay. The interaction effect is small but real, indicating a synergy layered on top of two already strong main predictors.

H2: High inpatient (>1) and High ED (GLM_OR = 0.66)

The Odds ratio is less than one and statistically significant in the opposite direction. This pattern likely reflects that the main effects already capture most of the risk signal. The appropriate course is to retain ED visits and prior admissions as separate predictors in the risk score and not rely on the "both" interaction.

H3: Abnormal A1C and Med change (GLM_OR = 0.98)

There is no evidence that the combination of an abnormal A1C (>7) and an in-hospital medication change creates additional readmission risk beyond their individual contributions. This was already seen from the bar diagram that patients who did not test A1C levels are among the highest readmitted patients. This is why we cannot confidently say that abnormal A1C result, even when paired with medication change does not increase readmission rate.

H4: AMA discharge vs Home (GLM_OR = 1.46)

Patients discharged against medical advice (AMA) have approximately **46%** higher odds of 30-day readmission than comparable patients discharged home, a strong and operationally meaningful result. This effect aligns with clinical experience that AMA discharges often reflect unresolved symptoms or limited readiness for discharge.

H5: High ED (≥3) and High Comorbidity (Q3) (GLM_OR = 0.75)

The interaction term is significantly below one, indicating no positive synergy beyond the main effects. If anything, the combined effect of very frequent ED use and high comorbidity is lower than expected from adding their separate influences. We should treat ED ≥3 and high comorbidity as separate risk factors in the predictive model and not depend on their interaction.

Interpretation of model performance:

- **Random Forest:** At first glance, the **Random Forest (RF)** model looks best with the **highest accuracy (0.829).** However, this is misleading. Our dataset is highly imbalanced, **89%** of the patients are not readmitted. The RF model's high accuracy comes from correctly predicting the easy **"No Readmission"** cases, but it fails to find the at-risk patients.

- **GLM:** This mode is the best at finding true readmissions with **highest Recall (0.491)** and best **AUC score of 0.675**, tied with **top F1 score of 0.279**. Lower precision reflects many false alarms, but it's the most useful for identifying patient at risk of readmission

- **Elastic Net:** This model performs identically to GLM (**Recall 0.486, AUC 0.675, F1 0.278**) with added stability. This a solid second choice when simpler model is preferred

- **XGBoost:** XGB Model performs moderately with **accuracy 0.776**, **recall 0.366**, **F1 0.269, AUC 0.651**. It is a better model than RF at finding positives but still trails GLM/ENet on the metrics that matter for intervention.

Recommended Model for Prediction

Based on this analysis, the **Generalized Linear Model (GLM) is the recommended model** for predicting 30-day hospital readmissions.

Justification:

1. **It finds the most at-risk patients:** It achieved the **highest Recall (0.491)**,
2. **It is the most reliable model:** It had the **highest AUC (0.675)**, proving it is the best model at distinguishing high-risk from low-risk patients.
3. **It has the best balance:** It also scored highest on the **F1-Score (0.279)**, showing it provides the best balance between finding patients and not creating too many false alarms.

Note: If the dataset were balanced, Random Forest and XGBoost would perform better in terms of overall accuracy. These models are well-suited when the main goal is general prediction accuracy rather than identifying rare readmission cases.

Recommendations for interventions or policy

The analysis identified clear patterns among patients most at risk of readmission within 30 days. These patients typically have longer hospital stays, take many medications, have multiple chronic conditions, or leave the hospital without proper discharge support. Based on our findings, particularly from supported hypotheses H1 and H4 and the top predictive factors, the following intervention plan is proposed.

1. Predictive Modeling Framework

The predictive model (GLM) should be embedded into the hospital's discharge system to automatically flag high-risk patients daily. Outcome tracking should occur monthly to monitor readmission rates and assess which interventions deliver the best results. Regular feedback sessions with care teams can refine the approach over time.

2. Support for AMA (Against Medical Advice) Discharges

Any patient initiating a discharge "Against Medical Advice" (AMA) is automatically flagged as **"High Risk."** This flag triggers an immediate consultation with a hospital social worker or case manager before the patient leaves. The goal is to understand and mitigate the patient's reasons for leaving (e.g., financial concerns, lack of home support). Regardless of outcome, the patient's file is sent to a post-discharge nurse team for an assertive follow-up call within 24 hours.

4. **Medication Reconciliation Plans for Complex Cases**
   Patients meeting the criteria for **H1 (Meds > 15 & Stay > 4 days) or H5 (ED ≥ 3 & High Comorbidity)** are flagged as **"Complex Cases"**. For these patients, a clinical pharmacist must conduct a full medication reconciliation and simplification review 48 hours before planned discharge. The goal is to reduce polypharmacy, eliminate redundant medications, and minimize the risk of adverse drug events post-discharge.

Limitations

- The Data set is imbalanced
- Coding quality for **discharge disposition** and **admission source** can bias H4/H1 interpretations.
- Interactions (especially H1) show **small** effect sizes; expect site-to-site variability.

Next steps

- **Pilot the AMA pathway** for **60–90** days; track readmissions and contact success.
- **Calibration & fairness:** reliability plots; subgroup AUC/Recall by sex/age/ethnicity.
- By balancing the data more accurate analysis can be done

Conclusion

This report successfully identified key, actionable drivers of 30-day hospital readmissions. Factors such as prior utilization, discharge status (especially AMA), and clinical instability (high medication count) are not just correlates but powerful, synergistic predictors of which patients will return.

Our predictive models demonstrate that it is possible to identify nearly half of all at-risk patients using a relatively simple, interpretable Generalized Linear Model. By deploying the targeted intervention strategies proposed, focused on case management for high utilizers, assertive follow-up for AMA patients, and specialized transitional care for unstable diabetic patients, a healthcare system can move from a reactive to a proactive stance on readmissions, improving patient outcomes and reducing systemic costs.