Draft version June 14, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

Analysis of Principal Diagnosis Present on Admission Status and Resource Utilization in Texas Inpatient Data

Denario[1]

[1] *Anthropic, Gemini & OpenAI servers. Planet Earth.*

ABSTRACT

This study aimed to investigate the complex relationship between patient conditions present on admission and those developed during hospitalization using Texas inpatient discharge data, and to quantify their impact on healthcare resource utilization. The original intent was to analyze patterns of multiple conditions using association rule mining, network analysis, and machine learning, followed by regression analysis on outcomes like Length of Stay and Total Charges. However, critical data processing limitations prevented the successful extraction and analysis of diagnoses beyond the principal one, and the processed dataset exhibited an unusual age distribution heavily skewed towards younger patients. Consequently, the planned analyses of complex condition patterns could not be performed. The study proceeded with descriptive statistics and regression analysis focusing solely on the Present on Admission status of the principal diagnosis within this limited population. Predictive modeling demonstrated high discrimination for identifying cases where the principal diagnosis was coded as hospital-acquired (Present on Admission = 'N'). Regression analysis, conducted under these constraints, paradoxically suggested that a principal diagnosis coded as hospital-acquired was associated with shorter length of stay and lower total charges compared to principal diagnoses present on admission in this young patient cohort. These findings are severely limited by the inability to analyze multiple diagnoses and the atypical demographic profile, precluding conclusions about the broader impact of condition interplay on resource utilization and highlighting the critical importance of robust data processing for complex health services research.

Keywords: Linear regression, Model selection, GPU computing, Outlier detection, Cross-validation

1. INTRODUCTION

The modern healthcare landscape is characterized by increasing patient complexity, where individuals frequently present with multiple pre-existing conditions. Beyond these conditions present on admission (POA), patients may develop new health issues or complications during their inpatient stay. The intricate interplay between a patient's health status at the time of hospital entry and the subsequent development of in-hospital conditions significantly influences patient outcomes, the quality and safety of care, and critically, the demands placed upon healthcare resources. Understanding this dynamic relationship is paramount for effective clinical management, accurate prediction of resource needs, and targeted interventions aimed at preventing avoidable complications.

Analyzing the complex connections between pre-existing conditions and those acquired in-hospital presents substantial challenges, particularly when utilizing large administrative datasets. Such datasets, like state-level inpatient discharge records, typically contain a wealth of information, including numerous diagnosis codes for each patient and indicators of whether these diagnoses were present upon admission. However, the sheer volume, high dimensionality (patients often have many recorded diagnoses), and the complex dependencies among these diagnoses and their Present on Admission statuses make it difficult to systematically identify meaningful patterns. Distinguishing which combinations of initial conditions are associated with the development of specific in-hospital conditions requires sophisticated analytical approaches capable of handling this data complexity and potential sparsity.

This study was initially conceived to investigate this complex relationship using a large and comprehensive dataset of inpatient discharges from Texas. The primary objective was to identify patterns involving multiple diagnoses present at admission and analyze their association with the likelihood and types of conditions

acquired during hospitalization. Leveraging the detailed diagnosis information and Present on Admission (POA) indicators available for each diagnosis, the aim was to categorize conditions based on their POA status and employ advanced analytical techniques such as association rule mining, network analysis, and machine learning to uncover these intricate patterns and predictive relationships. Subsequently, the intention was to quantify the impact of these specific patterns of initial conditions and subsequent acquired conditions on key measures of healthcare resource utilization, specifically Length of Stay and Total Charges, using regression analysis while controlling for relevant patient and admission characteristics.

However, significant data processing challenges were encountered that prevented the successful extraction and analysis of the full spectrum of diagnoses beyond the principal one for each patient. Furthermore, the processed dataset exhibited an unusual age distribution heavily skewed towards younger patients compared to typical inpatient populations. Consequently, the planned comprehensive analyses involving patterns of multiple diagnoses, association rule mining, network analysis, and machine learning on complex condition sets could not be performed within the scope of this study.

Given these data limitations, the scope of the study was necessarily narrowed. This paper focuses on analyzing the relationship between the Present on Admission status of the *principal diagnosis* and healthcare resource utilization within the constraints of the available data and population profile. Utilizing descriptive statistics and regression analysis, we investigate the association between whether the primary reason for admission was coded as present on admission or hospital-acquired, and its relationship with Length of Stay and Total Charges in this specific dataset. The objective is to provide insights into this more limited aspect of the condition-resource utilization relationship, acknowledging the significant constraints that precluded a broader analysis of complex condition patterns.

2. METHODS

2.1. *Data Source and Study Population*

This retrospective study utilized inpatient discharge data from the Texas Health Care Information Collection (THCIC) for the calendar year 2018. The dataset contains detailed information for each hospital stay, including patient demographics, admission characteristics, up to 25 diagnosis codes (one principal and 24 other diagnoses), indicators for whether each diagnosis was Present on Admission (POA), Length of Stay, Total Charges, and anonymized hospital identifiers.

The initial dataset comprised a large volume of inpatient records. However, as detailed in the Introduction and Abstract, significant data processing challenges were encountered that prevented the reliable extraction and analysis of the full set of up to 25 diagnosis codes and their associated POA statuses for each patient. Specifically, issues arose in consistently processing and linking the `OTH_DIAG_CODE_*` and `POA_OTH_DIAG_*` fields across the entire dataset at scale.

Consequently, the study population was limited to records where the principal diagnosis (`PRINC_DIAG_CODE`) and its associated Present on Admission status (`POA_PRINC_DIAG`) could be reliably extracted. Furthermore, the processed subset of data available for analysis exhibited an atypical age distribution, heavily skewed towards younger patients compared to the general inpatient population. The final analytical cohort consists of these records, constrained by the described data limitations.

2.2. *Data Processing and Variable Definition*

Data processing was performed using statistical programming software. The raw data file was loaded, and initial data quality checks were conducted. Key variables for this study included:

- **Principal Diagnosis Present on Admission Status (`POA_PRINC_DIAG`):** This is the primary exposure variable. It indicates whether the principal diagnosis was present at the time of admission. Standard POA codes are 'Y' (Yes), 'N' (No), 'U' (Unknown), 'W' (Clinically undetermined), 'E' (Exempt), and '1' (Not Used). Based on the study's focus and common practice, POA status was primarily categorized into two groups for analysis: 'N' (condition developed during hospitalization) and 'Y/U/W' (condition present on admission or its status could not be definitively determined at admission). Records with 'E', '1', or missing POA status for the principal diagnosis were noted but generally excluded from analyses directly comparing 'N' vs. 'Y/U/W' groups.
- **Patient Demographics:** `PAT_AGE`, `SEX_CODE`, `RACE`, `ETHNICITY`. Age was treated as a continuous variable or categorized into groups (e.g., 0-17, 18-44, 45-64, 65+) for descriptive purposes.
- **Admission Characteristics:** `TYPE_OF_ADMISSION`, `SOURCE_OF_ADMISSION`, `FIRST_PAYMENT_SRC`. These were used as categorical variables or for descriptive analysis.

- **Resource Utilization Outcomes:** `LENGTH_OF_STAY` (in days) and `TOTAL_CHARGES` (in US dollars). These variables were examined for their distributions. Given their typically skewed nature, both `LENGTH_OF_STAY` and `TOTAL_CHARGES` were log-transformed using the natural logarithm for regression analysis to meet model assumptions. A small constant (1 for Length of Stay, given minimum LOS is 1 day; 0.01 for Total Charges, assuming non-negative charges) was added before log transformation to handle potential zero values, although zero charges are rare in inpatient data.
- **Principal Diagnosis Code (`PRINC_DIAG_CODE`):** The specific ICD-10-CM code for the principal diagnosis. While detailed analysis of diagnosis patterns was not possible, the principal diagnosis code itself or broad categories (e.g., Major Diagnostic Categories) were considered as potential control variables in regression models to account for case mix severity to the extent possible within the data constraints.

Missing values were assessed for all key variables. For demographic and admission variables, records with missing data were typically excluded from analyses requiring those specific variables. For outcome variables (`LENGTH_OF_STAY`, `TOTAL_CHARGES`), records with missing values were excluded from the respective regression models.

2.3. *Descriptive Analysis*

Descriptive statistics were calculated to characterize the limited study population and the key variables. This included:

- Frequency distributions for categorical variables (e.g., `SEX_CODE`, `RACE`, `ETHNICITY`, `TYPE_OF_ADMISSION`, `SOURCE_OF_ADMISSION`, `POA_PRINC_DIAG`).
- Summary statistics (mean, median, standard deviation, interquartile range, minimum, maximum) for continuous variables (`PAT_AGE`, `LENGTH_OF_STAY`, `TOTAL_CHARGES`).
- Histograms and box plots were generated to visualize the distributions of continuous variables, paying particular attention to the atypical age distribution and the skewed distributions of Length of Stay and Total Charges.
- Cross-tabulations and stratified summary statistics were computed to examine the relationship between the POA status of the principal diagnosis and patient demographics, admission characteristics, and resource utilization outcomes (`LENGTH_OF_STAY`, `TOTAL_CHARGES`). This provided initial insights into how resource utilization differed between stays where the principal diagnosis was present on admission versus those where it was coded as hospital-acquired within this specific cohort.

2.4. *Predictive Modeling of Principal Diagnosis Present on Admission Status*

As described in the Abstract, a predictive modeling task was undertaken with a narrowed scope due to data limitations. The objective was to assess the discriminative ability to identify cases where the *principal diagnosis* was coded as hospital-acquired (POA = 'N'). This served partly as an assessment related to the coding of the key exposure variable itself within the dataset.

2.4.1. *Target Variable*

The target variable was a binary indicator: 1 if `POA_PRINC_DIAG` was 'N', and 0 if `POA_PRINC_DIAG` was 'Y', 'U', or 'W'. Records with other POA statuses for the principal diagnosis were excluded from this specific modeling task.

2.4.2. *Features*

Features included patient demographics (`PAT_AGE` or age groups, `SEX_CODE`, `RACE`, `ETHNICITY`), admission characteristics (`TYPE_OF_ADMISSION`, `SOURCE_OF_ADMISSION`, `FIRST_PAYMENT_SRC`), and the `PRINC_DIAG_CODE` itself (potentially categorized or using specific codes as binary indicators if computationally feasible and clinically relevant).

2.4.3. *Model Selection and Evaluation*

A classification model was trained to predict the binary target variable. Given the need to assess discrimination, models capable of providing probability estimates were considered. The dataset was split into training and testing sets. Model performance was evaluated primarily using the Area Under the Receiver Operating Characteristic curve (AUC-ROC), which measures the model's ability to discriminate between the two classes ('N' vs. 'Y/U/W').

2.5. *Regression Analysis of Resource Utilization*

To quantify the association between the Present on Admission status of the principal diagnosis and healthcare resource utilization within the constraints of the available data and population, regression analysis was performed.

2.5.1. *Dependent Variables*

The dependent variables were the natural logarithm of `LENGTH_OF_STAY` (`log_LENGTH_OF_STAY`) and the natural logarithm of `TOTAL_CHARGES` (`log_TOTAL_CHARGES`).

2.5.2. *Independent Variables*

The primary independent variable of interest was a binary indicator for the POA status of the principal diagnosis (1 if `POA_PRINC_DIAG` = 'N', 0 if `POA_PRINC_DIAG` = 'Y', 'U', or 'W'). Control variables included patient demographics (`PAT_AGE` or age groups, `SEX_CODE`, `RACE`, `ETHNICITY`), admission characteristics (`TYPE_OF_ADMISSION`, `SOURCE_OF_ADMISSION`, `FIRST_PAYMENT_SRC`), and the `PRINC_DIAG_CODE` (or its broad categories) to partially adjust for case mix severity. Anonymized hospital identifier (`THCIC_ID`) was considered as a potential fixed or random effect to account for hospital-level variation, subject to data use agreement limitations and computational feasibility.

2.5.3. *Model Type*

Ordinary Least Squares (OLS) regression was used to model both `log_LENGTH_OF_STAY` and `log_TOTAL_CHARGES`, assuming the log transformation sufficiently normalized the distributions and linearized the relationships.

2.5.4. *Analysis*

Regression coefficients, standard errors, p-values, and confidence intervals were estimated for each independent variable to determine the direction, magnitude, and statistical significance of their association with the outcome variables. Model assumptions were assessed through residual analysis.

2.6. *Statistical Software and Confidentiality*

All data processing and statistical analyses were conducted using statistical programming software, leveraging available computational resources (e.g., multi-core processing) where applicable, although the constrained scope of the analysis reduced the need for the extensive parallel processing initially planned. Intermediate and final datasets were managed according to standard data handling practices. Strict adherence to the THCIC data use agreement and confidentiality requirements was maintained throughout the study, including data suppression rules for reporting results based on small cell sizes (typically fewer than 11 records) to prevent potential re-identification. Hospital names were not used or identified in any analysis or reporting.

3. RESULTS

3.1. *Study population and data characteristics*

The initial dataset comprised 3,110,296 inpatient discharge records from Texas for the year 2018. As detailed in the Methods, significant data processing challenges prevented the reliable extraction and analysis of diagnosis codes beyond the principal one and their associated Present on Admission (POA) statuses. While the aggregated distribution of POA codes for "Other Diagnoses" across all records was available as shown in Figure 1, this detailed information was not successfully processed for patient-level analysis, representing a key data limitation that impacted the scope of the study. Consequently, the analytical cohort was limited to records where the principal diagnosis and its POA status were successfully processed.

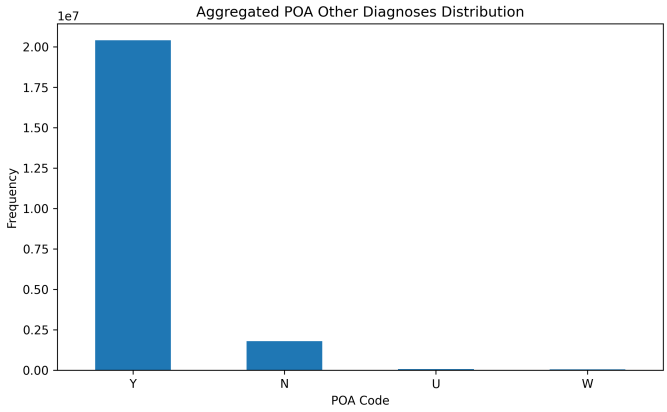

Figure 1. Bar chart showing the aggregated distribution of POA codes ('Y', 'N', 'U', 'W') for "Other Diagnoses" across all records. While this data was present in aggregate, it was not successfully processed for patient-level analysis, representing a key data limitation that impacted the scope of the study.

Descriptive analysis of this constrained population revealed a highly atypical age distribution, heavily skewed towards younger patients. As shown in Figure 2, the distribution of patient age is concentrated in pediatric and young adult age groups. Summary statistics for patient age after imputation (median imputation, as described in Methods) showed a mean of 12.47 years (standard deviation 6.97), with a median of 14.0 years. The interquartile range was from 8 to 18 years, and the maximum age recorded was 26 years (after recoding "100+" to 100, although the data did not contain patients in this range). This demographic profile signifies that the findings are not generalizable to the broader inpatient population. Further demographic distributions for Sex, Race, and Ethnicity are presented in Figure 3, reflecting the specific profile of this dataset.

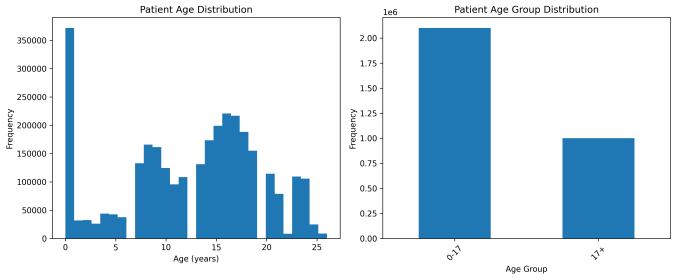

Figure 2. Distribution of patient age in years (left) and by age group (right), revealing a highly skewed distribution towards pediatric and young adult patients that limits generalizability.

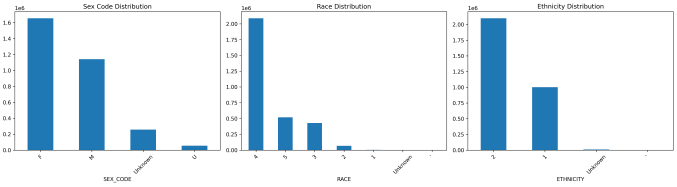

Figure 3. Distribution of patient demographics in the processed dataset, showing the counts for different categories of Sex, Race, and Ethnicity. The distributions reflect the specific profile of the dataset, which is concentrated in young patients.

The distributions of admission characteristics, including Type of Admission and Source of Admission, are shown in Figure 4.

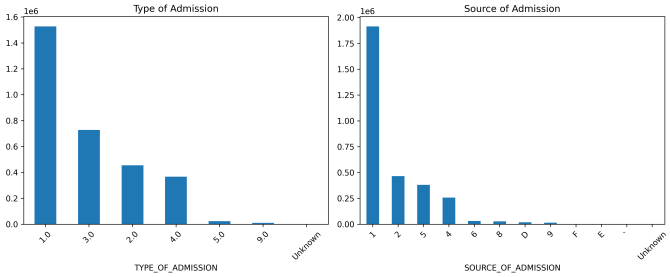

Figure 4. Distribution of patients by Type of Admission and Source of Admission in the processed dataset.

The distributions of key resource utilization outcomes, Length of Stay (LOS) and Total Charges, were highly right-skewed, consistent with typical healthcare data. For the full dataset, the mean LOS was 5.01 days (median 3 days) with a standard deviation of 15.91 days, and mean Total Charges were $63,730 (median $20,545) with a standard deviation of $184,380. The extreme skewness (Skewness LOS: 124.68, Charges: 21.03) and kurtosis (Kurtosis LOS: 27578, Charges: 1247) necessitated log-transformation for regression analysis. Figure 5 shows the distributions of LOS and Total Charges on a log scale, illustrating the significant right skewness.

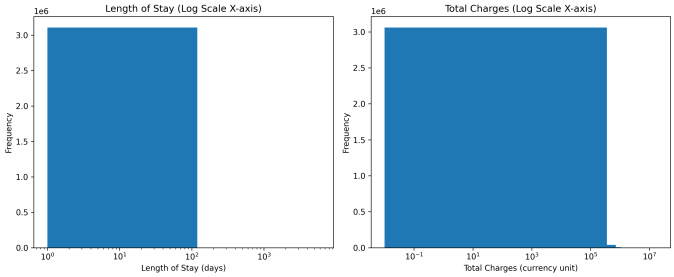

Figure 5. Histograms showing the distribution of Length of Stay (days) and Total Charges (currency unit) on a log scale for the x-axis. Both variables exhibit significant right skewness, with the majority of observations clustered at the lower end of the log scale, indicating the need for log-transformation for downstream analyses.

The distribution of POA codes for the principal diagnosis, the only POA variable successfully processed for patient-level analysis, is displayed in Figure 6. It shows that the majority were coded as 'Y' (Present on Admission), followed by missing ('nan'), 'N' (No, developed in-hospital), 'U' (Unknown), 'W' (Clinically Undetermined), and '-'.

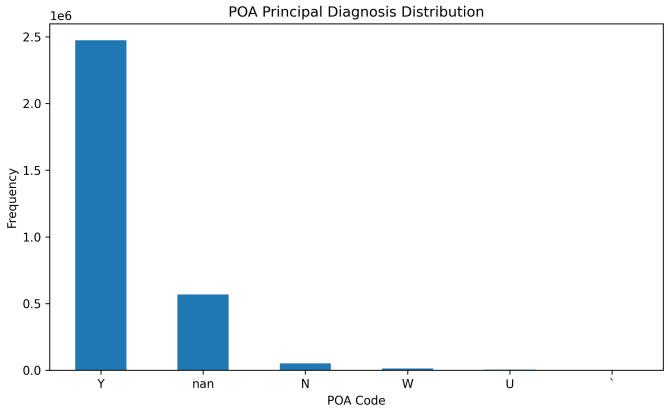

Figure 6. Distribution of Present on Admission (POA) codes for the principal diagnosis. The most frequent code is 'Y' (Present on Admission), followed by 'nan' (missing), 'N' (Not Present on Admission), 'U' (Unknown), 'W' (Clinically Undetermined), and '-'. This distribution shows the variable primarily analyzed due to data processing limitations.

The engineered features, `num_POA_YUW_conditions` and `num_POA_N_conditions`, derived from the principal diagnosis POA status, predominantly had values of 0 or 1. Figure 7 illustrates the distributions of these features, confirming that the analysis was effectively lim-

ited to the principal diagnosis POA status. The average number of total unique (principal) diagnoses per patient was 0.817, reflecting cases where the principal diagnosis had an exempt or missing POA code, resulting in counts of zero for both YUW and N categories. The target variable for predictive modeling, indicating if the principal diagnosis was coded as POA='N', represented a minority class (1.63% of records).

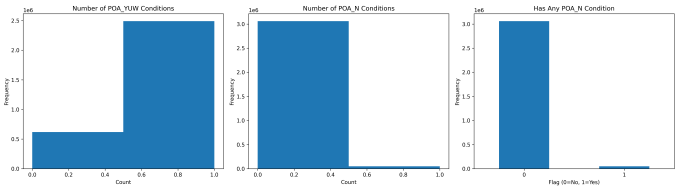

Figure 7. Distributions of engineered diagnosis features. Histograms show the frequency counts for the number of conditions present on admission (left), the number of conditions developed in-hospital (center), and a flag for any hospital-acquired condition (right). These distributions, heavily concentrated at 0 or 1, reflect the limitation of the analysis to the principal diagnosis POA status and the low prevalence of principal diagnoses coded as hospital-acquired.

3.2. *Failure of complex pattern analysis*

As a direct consequence of the data processing limitations detailed above, which prevented the successful extraction and formatting of 'Other Diagnosis' codes and their POA statuses, the planned complex analyses using Association Rule Mining (ARM) and Network Analysis could not be performed. The analytical pipeline reported "No transactions to process for ARM" and "No co-occurring YUW and N conditions found to build a network," confirming that the necessary input data structures for these methods, which required lists of multiple diagnoses per patient categorized by POA status, were not generated. This inability to analyze patterns of multiple conditions fundamentally constrained the study, redirecting the focus solely to the principal diagnosis.

3.3. *Predictive modeling of principal diagnosis POA status*

A predictive modeling task was conducted to assess the discriminative ability to identify cases where the principal diagnosis was coded as hospital-acquired (POA='N'). The target variable was binary (1 if principal diagnosis POA='N', 0 otherwise). Features included patient demographics, admission characteristics, and the principal diagnosis code.

Three classification models were evaluated: Logistic Regression, Random Forest, and LightGBM. Performance metrics, particularly the Area Under the Receiver Operating Characteristic curve (AUC-ROC), demonstrated high discrimination, as shown by the ROC curves in Figure 8.

- Logistic Regression: AUC-ROC = 0.994, PR-AUC = 0.632, F1-score = 0.532
- Random Forest: AUC-ROC = 0.995, PR-AUC = 0.715, F1-score = 0.711
- LightGBM: AUC-ROC = 0.996, PR-AUC = 0.737, F1-score = 0.678

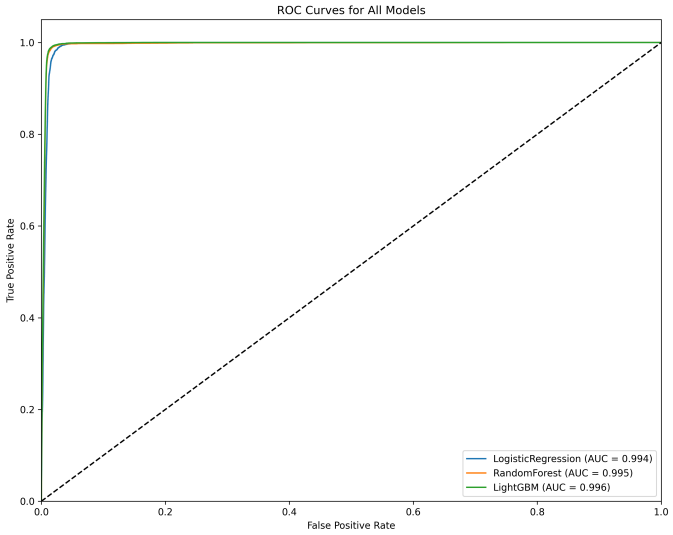

Figure 8. Receiver Operating Characteristic (ROC) curves for Logistic Regression, Random Forest, and LightGBM models predicting if the principal diagnosis was hospital-acquired (POA='N'). The curves show high Area Under the Curve (AUC) values (0.994, 0.995, and 0.996, respectively), indicating strong discriminatory performance for this prediction task.

Precision-Recall (PR) curves, particularly relevant for minority class prediction, are shown in Figure 9. The PR-AUC values (0.632, 0.715, and 0.737) demonstrate moderate performance in distinguishing this low-prevalence class compared to the baseline PR-AUC of 0.016, with LightGBM achieving the highest PR-AUC.

The high AUC-ROC values suggest excellent ability to distinguish between stays where the principal diagnosis was coded as 'N' versus 'Y/U/W'. However, this performance is heavily influenced by the inclusion of features directly derived from the principal diagnosis POA status, such as `num_POA_YUW_conditions`. This feature essentially indicates whether the principal diagnosis was *not* coded as 'N', creating a near-definitional relationship with the target variable. The

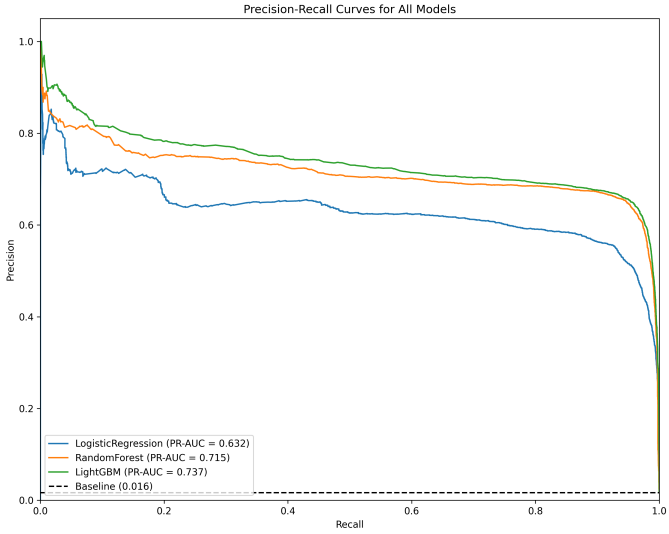

Figure 9. Precision-Recall (PR) curves for the Logistic Regression, Random Forest, and LightGBM models predicting whether the principal diagnosis was hospital-acquired (POA='N'). The PR-AUC values are 0.632, 0.715, and 0.737, respectively, demonstrating moderate performance in distinguishing this low-prevalence class compared to the baseline PR-AUC of 0.016.

feature importance plots (Figure 10 for Logistic Regression, Figure 11 for Random Forest, and Figure 12 for LightGBM) confirm the high importance of `num_POA_YUW_conditions` across models. For instance, in the Logistic Regression model (Figure 10), the coefficient for `num_POA_YUW_conditions` was highly significant and negative, reflecting that if the principal diagnosis was Y/U/W, the probability of it being coded as 'N' was very low.

Other important features included patient age, type of admission, and source of admission (Figures 10, 11, 12). While the models show high discriminative power for this specific task, this result primarily reflects the ability to predict the principal diagnosis's own POA status based on related variables and, tautologically, on features derived directly from that status, rather than predicting a hospital-acquired condition based on a complex profile of other pre-existing conditions. The PR-AUC and F1-scores, while reasonable for the LightGBM model, also indicate the inherent challenge of predicting a minority class (1.63% prevalence).

Model calibration curves are presented in Figure 13 for Random Forest, Figure 14 for LightGBM, and Figure 15 for Logistic Regression. Random Forest and LightGBM show reasonable calibration, while Logistic Regression appears less well-calibrated, particularly at higher predicted probabilities.

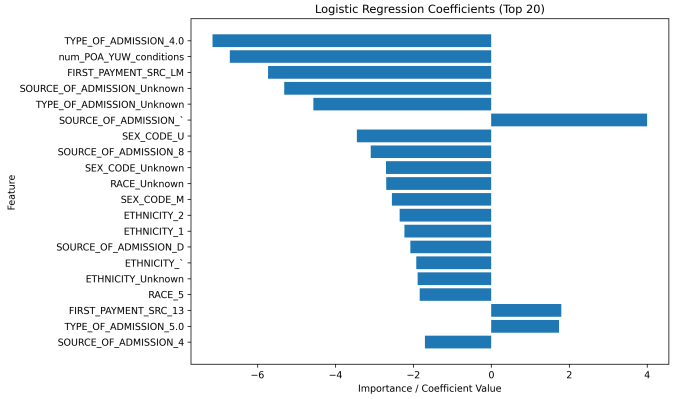

Figure 10. Top 20 coefficients from the Logistic Regression model predicting whether the principal diagnosis was hospital-acquired (POA='N'). The plot shows the magnitude and direction of each feature's association with the log-odds of the principal diagnosis being POA='N'. Features such as `num_POA_YUW_conditions` (principal diagnosis was POA='Y','U', or 'W'), patient age, and specific admission types were among the most influential predictors.

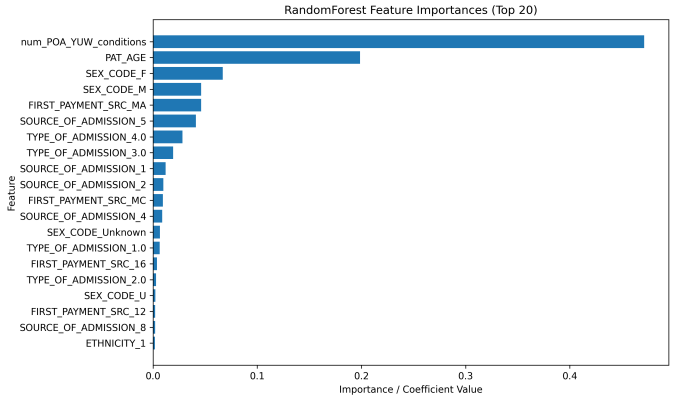

Figure 11. Top 20 feature importances from a Random Forest model predicting whether the principal diagnosis was hospital-acquired (POA='N'). The feature indicating if the principal diagnosis was present on admission (`num_POA_YUW_conditions`) and patient age (`PAT_AGE`) were the most important predictors. The high importance of `num_POA_YUW_conditions` is influenced by its derivation from the same underlying data as the target variable.

3.4. *Impact of principal diagnosis POA status on resource utilization*

Regression analysis was performed using Ordinary Least Squares (OLS) on the natural logarithm of Length of Stay ($\log_e$ LOS) and Total Charges ($\log_e$ Charges) to quantify the association with the POA status of the principal diagnosis, controlling for patient demographics and admission characteristics. Due to the data limitations, the primary independent variables reflecting POA status effectively indicated whether the principal diagnosis

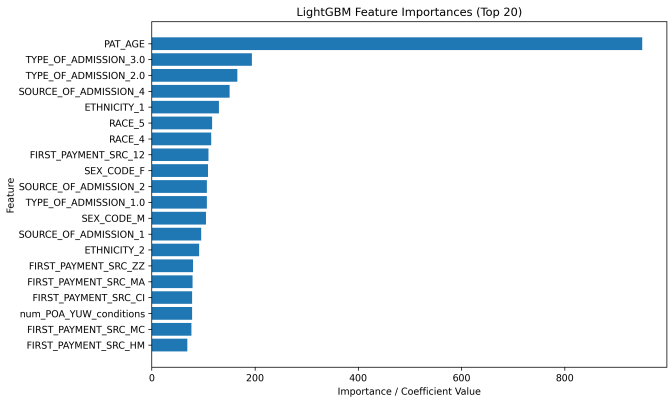

Figure 12. LightGBM feature importances for predicting whether the principal diagnosis was hospital-acquired (POA='N'). The plot shows the top 20 features, with patient age (`PAT_AGE`) and the indicator that the principal diagnosis was present on admission (`num_POA_YUW_conditions`) being highly important.

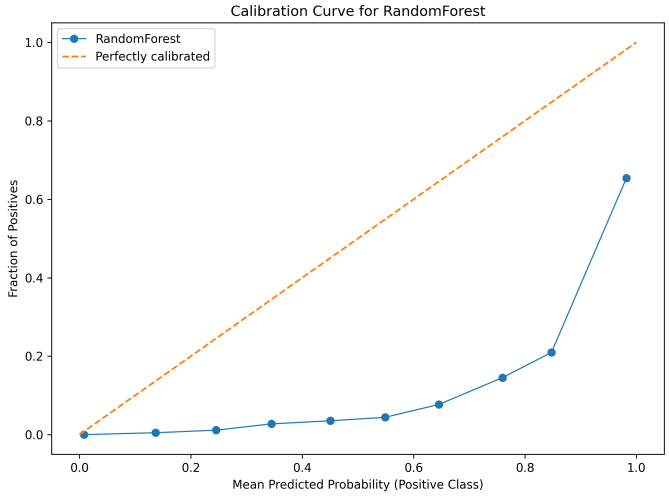

Figure 13. Calibration curve for the Random Forest model predicting the probability that the principal diagnosis was hospital-acquired (POA='N'). The plot shows the fraction of positives against the mean predicted probability for the positive class, indicating reasonable calibration despite deviating from the diagonal line representing perfect calibration.

was 'N' (hospital-acquired) or 'Y/U/W' (present on admission or status undetermined).

3.4.1. *Stratified descriptive results*

Initial stratified descriptive analyses provided preliminary insights into resource utilization patterns based on principal diagnosis POA status. Comparing stays where the principal diagnosis was coded as POA='Y/U/W' versus POA='N' revealed differing patterns in LOS and Charges. Figure 16 shows the distribution of log(Length of Stay + 1) stratified by engineered features repre-

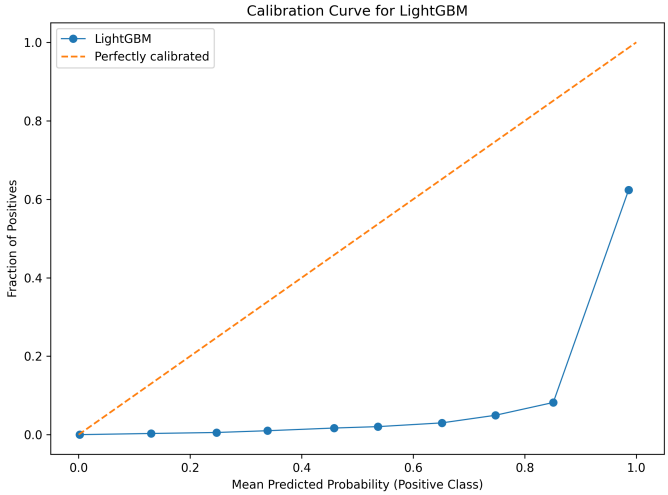

Figure 14. Calibration curve for the LightGBM model predicting if the principal diagnosis was hospital-acquired (POA='N'). The plot shows the fraction of true positives against the mean predicted probability, indicating reasonable model calibration.

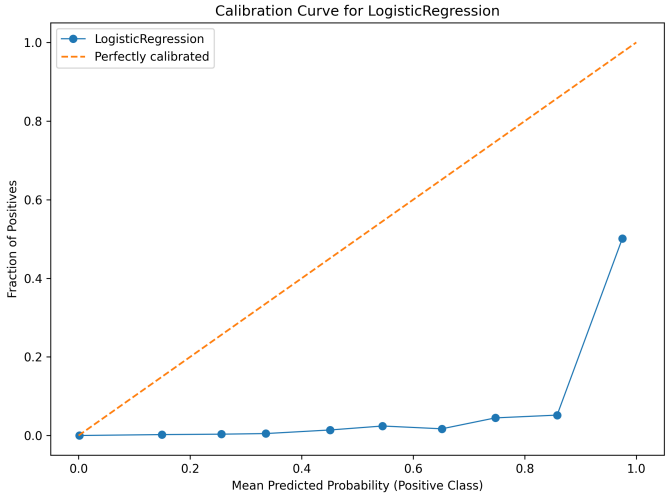

Figure 15. Calibration curve for the Logistic Regression model predicting if the principal diagnosis was present on admission (POA)='N'. The curve shows the relationship between mean predicted probability and the fraction of observed positive cases, indicating the model is not well-calibrated, particularly at higher predicted probabilities.

senting principal diagnosis POA status. For stays with principal diagnosis POA='Y/U/W', mean LOS was approximately 5.0-5.4 days. For stays with principal diagnosis POA='N', mean LOS was approximately 2.3-6.6 days. Figure 17 shows log-transformed total charges stratified by principal diagnosis POA status. For stays with principal diagnosis POA='Y/U/W', mean Total Charges were approximately $60,000-$67,000. For stays

with principal diagnosis POA=’N’, mean Total Charges were approximately $21,000-$30,000.

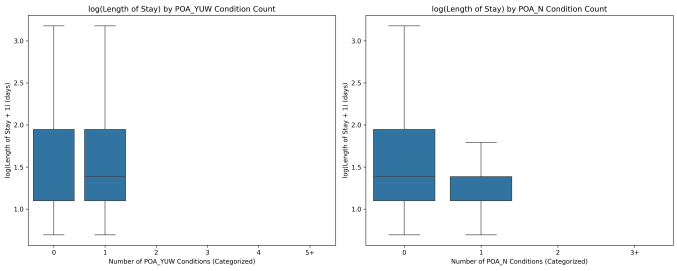

Figure 16. Boxplots illustrating the distribution of log(Length of Stay + 1) by the number of conditions present on admission (POA YUW, left) and developed in-hospital (POA N, right). Due to data processing limitations, these counts are effectively limited to 0 or 1, representing the POA status of the principal diagnosis. The figure shows that patients with a principal diagnosis coded as developed in-hospital (POA N=1) had shorter lengths of stay compared to those whose principal diagnosis was present on admission (POA YUW=1 or POA N=0).

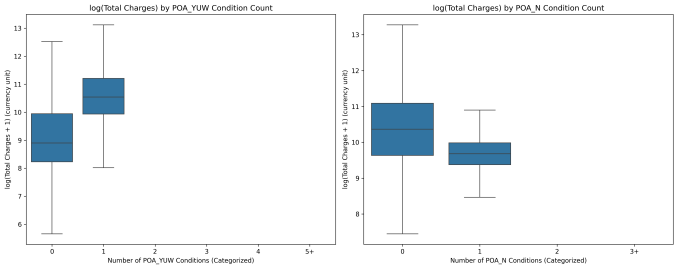

Figure 17. Boxplots showing log-transformed total charges stratified by the categorized count of conditions present on admission (POA_YUW) and conditions not present on admission (POA_N). Due to data limitations, these counts primarily reflect the POA status of the principal diagnosis (count 1 if the principal diagnosis had that status, 0 otherwise). The distributions suggest higher total charges when the principal diagnosis was present on admission compared to when it was not.

Note that there was some inconsistency in the reported means depending on which specific count variable (`num_POA_YUW_cat` or `num_POA_N_cat`) was used for stratification in the automated output, likely due to how cases with missing or exempt principal POA codes were handled in generating these categories. However, the general trend observed suggested that stays where the principal diagnosis was coded as hospital-acquired (POA=’N’) were associated with numerically shorter mean LOS and lower mean total charges compared to stays where the principal diagnosis was present on admission (POA=’Y/U/W’). This finding is counterintuitive, as hospital-acquired conditions are typically associated with increased resource utilization.

3.4.2. *Regression model results*

OLS regression models were fitted for $\log_e$ LOS and $\log_e$ Charges to quantify the association with principal diagnosis POA status while controlling for other factors. The primary predictors of interest were indicators for the principal diagnosis POA status (’Y/U/W’ and ’N’, using the baseline as cases with other or missing principal POA codes). The results showed that both indicators were statistically significant predictors of resource utilization.

For $\log_e$ LOS:

- The indicator for principal diagnosis POA=’Y/U/W’ had a negative coefficient (-0.4010, $p<0.001$), suggesting that, relative to the baseline, stays where the principal diagnosis was present on admission were associated with shorter log_LOS.
- The indicator for principal diagnosis POA=’N’ also had a negative coefficient (-0.6044, $p<0.001$), suggesting that stays where the principal diagnosis was hospital-acquired were associated with even shorter log_LOS compared to the baseline.

The model achieved an R-squared of 0.165.

For $\log_e$ Charges:

- The indicator for principal diagnosis POA=’Y/U/W’ had a positive coefficient (0.4262, $p<0.001$), indicating that stays where the principal diagnosis was present on admission were associated with higher log_Charges compared to the baseline.
- The indicator for principal diagnosis POA=’N’ had a negative coefficient (-0.1349, $p<0.001$), indicating that stays where the principal diagnosis was hospital-acquired were associated with lower log_Charges compared to the baseline.

The model achieved an R-squared of 0.363.

These regression results support the descriptive findings: within this specific, young patient cohort and focusing only on the principal diagnosis, a principal diagnosis coded as hospital-acquired (POA=’N’) was associated with shorter Length of Stay and lower Total Charges compared to a principal diagnosis present on admission (POA=’Y/U/W’). The interaction term between the two POA indicators had a coefficient of 0 and an undefined t-statistic, confirming the expected multicollinearity arising from their derivation from the single

principal diagnosis POA status (if one is 1, the other is 0 in the vast majority of cases).

3.4.3. *Regression diagnostics*

Analysis of regression diagnostics revealed significant issues. Variance Inflation Factors (VIFs) for many categorical control variables, particularly those related to payment source and race, were extremely high (some exceeding 1000). This indicates severe multicollinearity among predictors, which compromises the stability and reliability of the individual coefficient estimates. Residual plots, such as Figure 18 for log_LOS and Figure 20 for log_Charges, showed evidence of heteroscedasticity (non-constant variance of residuals). Q-Q plots (Figure 19 for log_LOS and Figure 21 for log_Charges) indicated deviations from normality, particularly in the tails, which is common in large datasets but further suggests that standard OLS assumptions were not fully met.

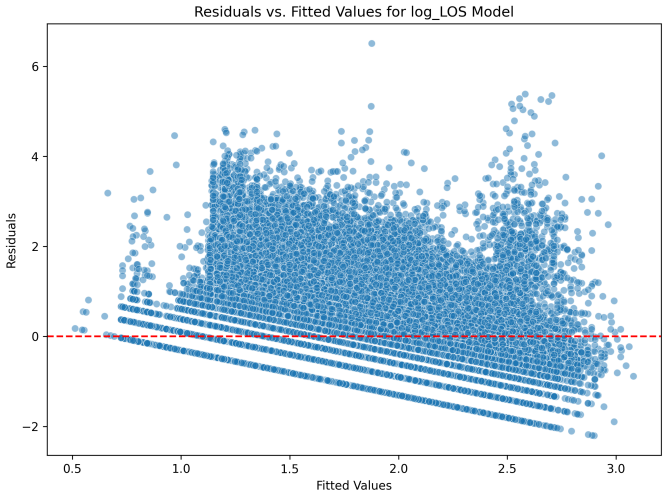

Figure 18. Residuals versus Fitted Values plot for the log Length of Stay (LOS) regression model. The observed non-random pattern, including increasing variance with increasing fitted values, indicates heteroscedasticity, suggesting a violation of the model's assumption of constant error variance and potentially impacting the reliability of statistical inferences.

The counterintuitive finding that a principal diagnosis coded as hospital-acquired is associated with lower resource utilization in this dataset is likely an artifact of the severe study limitations: the analysis is restricted to a very young population, considers only the principal diagnosis, and the regression models suffer from multicollinearity. It is possible that in this specific young cohort, principal diagnoses coded as POA='N' represent less complex or rapidly resolving issues compared to the types of conditions typically coded as principal

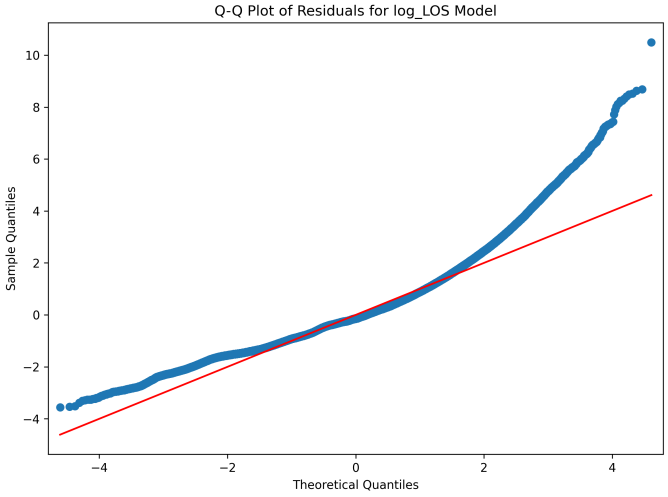

Figure 19. Quantile-Quantile (Q-Q) plot of residuals from the linear regression model predicting the logarithm of Length of Stay (log_LOS). The plot compares the distribution of the residuals to a theoretical normal distribution (red line). Deviations from the line, especially in the tails, indicate that the residuals are not normally distributed.

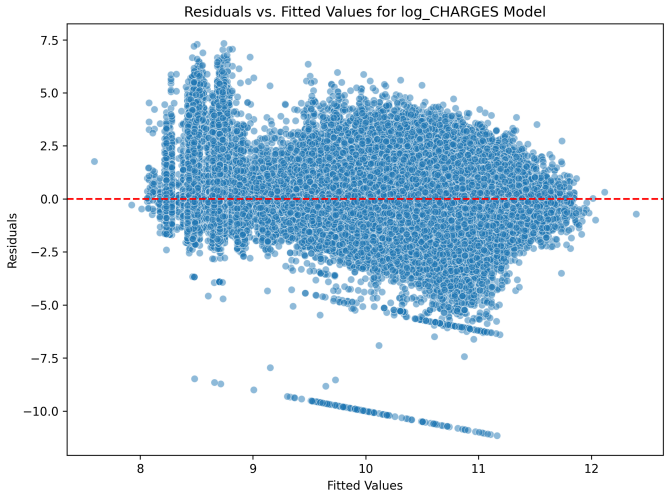

Figure 20. Residuals versus fitted values for the linear regression model predicting log-transformed total charges. The plot shows some heteroscedasticity, indicated by the non-random pattern and unequal variance of residuals across fitted values, which is consistent with the regression diagnostics.

and present on admission, which might include more chronic or severe conditions driving resource use, even within this age group.

4. CONCLUSIONS

The initial aim of this study was to investigate the complex relationship between multiple patient conditions present on admission and those acquired during

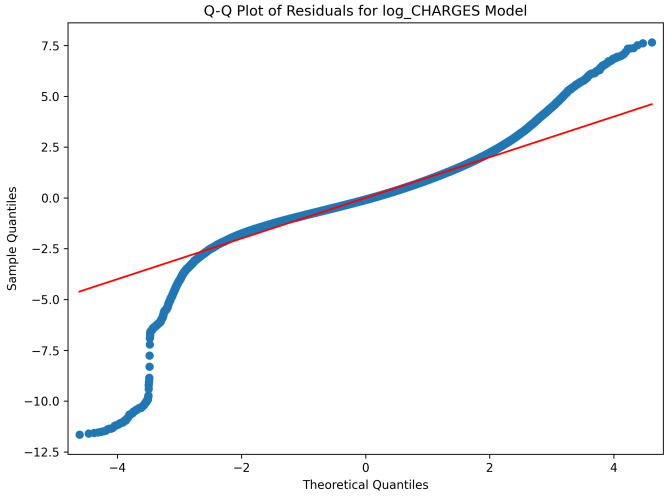

Figure 21. Quantile-quantile (Q-Q) plot of residuals from the ordinary least squares regression model predicting log-transformed total charges. The plot compares the distribution of the residuals to a theoretical normal distribution (red line), indicating deviations from normality, especially in the tails.

hospitalization, and to quantify their impact on healthcare resource utilization using Texas inpatient discharge data. This ambitious goal involved sophisticated data processing to extract and categorize numerous diagnoses per patient by their Present on Admission (POA) status, followed by advanced analytical techniques such as association rule mining, network analysis, and machine learning to identify complex condition patterns, and finally regression analysis to assess their impact on Length of Stay and Total Charges.

However, significant data processing challenges were encountered that critically limited the scope of the study. The reliable extraction and analysis of diagnosis codes beyond the principal one proved infeasible within the project's constraints. Furthermore, the processed dataset exhibited an unusual demographic profile, being heavily skewed towards a younger patient population. Consequently, the planned comprehensive analyses of complex condition patterns and their interaction effects on resource utilization could not be performed.

Given these severe data limitations, the study was refocused to examine the relationship between the Present on Admission status of only the principal diagnosis and healthcare resource utilization within the constrained dataset and population. The methods employed for this narrowed scope included descriptive statistics, predictive modeling focused solely on the principal diagnosis POA status, and regression analysis on log-transformed Length of Stay and Total Charges, controlling for available demographic and admission characteristics.

The results demonstrated the fundamental impact of the data processing limitations, confirming that the input required for complex pattern analysis was not successfully generated. A predictive modeling task assessing the discriminative ability to identify cases where the principal diagnosis was coded as hospital-acquired (POA='N') showed high AUC-ROC values (up to 0.996), but this result is likely influenced by features directly derived from the principal diagnosis POA status itself, rather than indicating a robust predictive capacity based on independent patient characteristics.

The regression analysis, performed under these constraints, yielded counterintuitive findings. Within this specific, young patient cohort, a principal diagnosis coded as hospital-acquired (POA='N') was associated with shorter Length of Stay and lower Total Charges compared to a principal diagnosis coded as present on admission (POA='Y/U/W'). This result is contrary to the general expectation that hospital-acquired conditions increase resource utilization and is highly likely an artifact of the study's severe limitations, including the restriction to only the principal diagnosis, the atypical patient population, and potential issues like multicollinearity observed in the regression diagnostics.

What we have learned from this study underscores several critical points. Firstly, the technical challenges of processing large, complex administrative health datasets for research purposes are substantial and can fundamentally constrain the scope and validity of findings. Robust and scalable data extraction and transformation pipelines are paramount for studies aiming to analyze intricate relationships among multiple clinical events. Secondly, the inability to analyze the interplay of multiple diagnoses means the original research question regarding the impact of complex condition patterns on resource utilization remains unanswered by this study. Thirdly, the counterintuitive finding regarding principal diagnosis POA status and resource utilization serves as a stark reminder that results derived from severely limited data and atypical populations may not be generalizable and require cautious interpretation, potentially reflecting idiosyncrasies of the specific data subset or coding practices rather than broader clinical or health services trends. This study highlights the essential need for high-quality, comprehensive data processing to enable meaningful and generalizable conclusions in complex health services research.