Draft version June 15, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

Modeling Inpatient Morbidity Dynamics Using Present on Admission Data: Predicting Emergent Conditions and Analyzing Resource Utilization in Texas Hospitals

Denario[1]

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

ABSTRACT

Understanding the dynamic evolution of patient health status during hospitalization is crucial for predicting outcomes and managing healthcare resources, yet traditional approaches often focus on static admission data. This study aimed to model inpatient morbidity dynamics by predicting the emergence of new conditions during hospitalization, defined using Present on Admission (POA) indicators, and quantifying their incremental impact on Length of Stay and Total Charges. We analyzed over 3.1 million inpatient discharge records from the 2018 Texas Hospital Inpatient Discharge data. Initial patient state was characterized by POA='Y' diagnoses, while emergent conditions were defined as POA='N' diagnoses. We employed machine learning models (Logistic Regression, Random Forest, XGBoost) to predict the likelihood of developing any emergent condition based on initial patient profiles and used regression models (Linear Regression, Random Forest, XGBoost) to assess the impact of emergent conditions on resource utilization, comparing models with and without emergent condition features, while also exploring variations across demographic subgroups and hospitals under strict confidentiality rules. Emergent conditions, as defined by POA='N', were identified in 1.63% of records. Models predicting the occurrence of any emergent condition achieved perfect or near-perfect classification scores, indicating a significant methodological issue, likely data leakage or a circular definition in feature engineering, which invalidates direct interpretation of these specific prediction results. For resource utilization, models explained up to 32% of the variance in Length of Stay and 57% in Log-Total Charges using initial patient characteristics. However, the inclusion of simple features indicating the presence or count of emergent conditions did not substantially improve predictive performance for either outcome when controlling for the initial patient profile. This study demonstrates the potential of using POA data to characterize dynamic morbidity but highlights critical challenges in accurately predicting the emergence of new conditions with the current approach, necessitating a re-evaluation of the prediction task formulation. Furthermore, within this framework, the simple occurrence of an emergent condition did not provide significant incremental explanatory power for resource utilization beyond the information available at admission, suggesting the need for more granular definitions of emergent morbidity or alternative modeling strategies to capture their true impact.

Keywords: Random Forests, Computational astronomy, Cross-validation, Computational methods, Linear regression

1. INTRODUCTION

Hospitalization represents a critical and often dynamic period in a patient's health trajectory. During an inpatient stay, a patient's clinical status can evolve significantly, influenced by their initial condition, the care received, and the development of new medical issues. Understanding and modeling this dynamic evolution of patient morbidity is fundamental for improving clinical decision-making, optimizing healthcare resource allocation, and ultimately enhancing patient safety and outcomes.

Traditional approaches to predicting hospital outcomes, such as length of stay or mortality, often rely on data captured primarily at the point of admission. While valuable, this static perspective fails to capture the complex interplay of factors and events that occur *during* the hospitalization itself. A significant contributor to changes in patient status, prolonged stays, increased costs, and adverse outcomes is the development of new conditions or complications during hospi-

talization, sometimes referred to as emergent morbidity or hospital-acquired conditions. Effectively identifying, predicting, and understanding the impact of these emergent conditions is a major challenge in healthcare analytics.

Modeling these dynamic changes and emergent events is inherently difficult. Biological processes are complex and non-linear, patient responses to illness and treatment vary widely, and hospital-specific care processes and environments can play a role. Furthermore, while healthcare administrative data is abundant, it often lacks the fine-grained temporal resolution needed to precisely track the onset of new conditions within a stay. A key challenge when using administrative data is distinguishing between conditions that genuinely emerge *during* the hospitalization and those that were present at admission but perhaps not fully documented until later in the stay.

This study seeks to address these challenges by leveraging the Present on Admission (POA) indicator, a data element available in many administrative discharge datasets. The POA indicator provides a crucial distinction, allowing us to identify diagnoses documented as present at the time of inpatient admission (typically coded as POA='Y') versus those that were not (coded as POA='N', 'U', 'W', 'E', or similar). By defining a patient's initial clinical state based on POA='Y' diagnoses and identifying emergent conditions based on POA='N' diagnoses, we can characterize a key aspect of the dynamic shift in patient morbidity during hospitalization using routinely collected data.

Specifically, this research utilizes a large dataset of over 3.1 million inpatient discharge records from Texas hospitals from 2018 to pursue two primary objectives. First, we aim to model the likelihood of developing any new condition (defined by the presence of at least one POA='N' diagnosis) based on the patient's characteristics and initial diagnosis profile at admission. This involves employing machine learning techniques to explore the predictability of emergent morbidity. Second, we seek to quantify the incremental impact of these emergent conditions on key healthcare resource utilization metrics: Length of Stay and Total Charges. We analyze the association between the presence and number of emergent conditions and these outcomes using regression models, controlling for the patient's initial profile to isolate the effect of the dynamic change. By comparing the performance and insights from models that include features related to emergent conditions against those that rely solely on admission data, we attempt to quantify the added explanatory power provided by capturing this dynamic aspect of morbidity. Furthermore, we explore variations in these dynamics and their impact across different patient demographic subgroups and reporting hospitals, while strictly adhering to data confidentiality requirements.

Through this approach, we move beyond a static view of inpatient care to gain insights into the process of morbidity evolution and its resource implications within a large, real-world hospital system, demonstrating the potential and inherent challenges of using POA data to characterize dynamic inpatient health states.

2. METHODS

This study utilized a retrospective cross-sectional design analyzing a large dataset of inpatient hospital discharges to model the dynamics of patient morbidity during hospitalization and assess their impact on resource utilization. The analytical approach involved several stages: data preparation and feature engineering to define initial and emergent morbidity using Present on Admission (POA) indicators, modeling the likelihood of developing emergent conditions, and regression analysis to quantify the association between emergent conditions and healthcare resource utilization metrics.

2.1. *Data Source and Study Population*

The primary data source for this study was the 2018 Texas Hospital Inpatient Discharge Public Use Data File (PUDF), provided by the Texas Health Care Information Collection (THCIC). This dataset comprises detailed information for over 3.1 million inpatient stays in Texas hospitals during the calendar year 2018. Each record represents a single hospital discharge and includes patient demographics, admission and discharge status, diagnosis and procedure codes with associated Present on Admission (POA) indicators, length of stay, total charges, and hospital identifiers. The analysis included all records present in the 2018 PUDF that passed the initial data cleaning and validation steps described below.

2.2. *Data Preparation and Feature Engineering*

Data preparation involved loading the raw data, systematic handling of missing values, correction of specific data fields, standardization of data types, and management of outliers. This phase built upon preliminary exploratory data analysis findings.

2.2.1. *Initial Cleaning and Missing Data Handling*

Datasets were loaded using pandas, ensuring appropriate memory management for large files. Columns identified as entirely empty or having extremely high rates of missingness (e.g., `Unnamed: 167`, `SPEC_UNIT_2`,

SPEC_UNIT_3, SPEC_UNIT_4, SPEC_UNIT_5) were removed. A systematic approach was applied to handle missing data in key fields:

- **Demographics:** Missing values and specific codes indicating unknown values ('U' for SEX_CODE, '`' for RACE and ETHNICITY) were mapped to an 'Unknown/Missing' category. Missing PAT_ZIP values were also grouped into an 'Unknown/Missing ZIP' category, handled cautiously due to confidentiality implications.
- **POA Indicators:** For all diagnosis codes (POA_PRINC_DIAG_CODE, POA_OTH_DIAG_X for $X = 1$ to 24), codes 'U' (Documentation insufficient), 'W' (Clinically undetermined), 'E' (Exempt from reporting), and any erroneous '1' or '‘' characters were mapped to a unified 'Non-Informative POA' category. Explicitly missing POA data (NaN) were treated as a distinct 'Missing POA Data' category, acknowledging the ambiguity regarding whether the condition was present on admission or developed later.
- **Administrative/Financial:** Missing values in SECONDARY_PAYMENT_SRC were explicitly categorized as 'No Secondary Payer'. For fields with very low missingness rates (TYPE_OF_ADMISSION, SOURCE_OF_ADMISSION, PAT_STATUS, FIRST_PAYMENT_SRC), missing values were either treated as a separate 'Unknown' category or, if counts were negligible, records with missing values in these specific fields were excluded from analyses requiring them.
- **Other Fields:** Missingness in fields like OPERATING_PHYSICIAN_UNIF_ID (44.0

2.2.2. *Patient Age Correction*

As indicated by the exploratory data analysis, the raw PAT_AGE field did not represent chronological age directly but rather a coded system or age bands. Based on external THCIC documentation, these codes were mapped to numeric age values. For age bands, the midpoint was used (e.g., 1-4 years mapped to 2.5). For the '<1 year' code, 0.5 years was assigned. For the '90+' code, 90 years was assigned. Any non-standard codes like '‘' were mapped to missing before conversion. A new column, PAT_AGE_NUMERIC_CORRECTED, was created to store the corrected numeric age.

2.2.3. *Data Type Conversion and Outlier Management*

LENGTH_OF_STAY was converted to an integer, and TOTAL_CHARGES to a float. Diagnosis, procedure, and E-codes were treated as string objects and standardized by removing whitespace and converting to uppercase. Categorical variables like SEX_CODE, RACE, ETHNICITY, PAT_STATUS, TYPE_OF_ADMISSION, SOURCE_OF_ADMISSION, FIRST_PAYMENT_SRC, cleaned PAT_ZIP, THCIC_ID, and DISCHARGE day of week were ensured to be consistently typed.

Outlier management was applied to key outcome variables:

- **Length of Stay:** While minimum LOS was 1 day, the maximum value was extremely high (5,905 days). Values exceeding 365 days were considered outliers and winsorized (capped) at 365 days to mitigate their influence, representing stays exceeding a typical acute care duration.
- **Total Charges:** Records with negative charges were set to NaN. Records with zero charges were reviewed; if deemed erroneous, they were set to a small positive value (e.g., $1) before transformation. For modeling, TOTAL_CHARGES was log-transformed using LOG_TOTAL_CHARGES = log(TOTAL_CHARGES + 1) to address skewness and handle zero/near-zero values. Extreme high values in log-transformed charges were examined but not winsorized unless they showed clear indication of data error after review.

2.2.4. *Feature Engineering for Morbidity Dynamics*

To capture the initial patient state and the emergence of new conditions, diagnosis codes were mapped to a higher-level grouping system, the Clinical Classifications Software Refined (CCSR), to reduce dimensionality and improve interpretability. This mapping was applied to all diagnosis codes.

Key features were engineered to define initial and emergent morbidity:

- **Initial State:** Diagnoses with a POA indicator of 'Y' were considered present on admission. Features included a list of raw and grouped initial diagnosis codes (INITIAL_DIAGNOSES_RAW, INITIAL_DIAGNOSES_GROUPED) and the count of initial diagnoses (NUM_INITIAL_DIAGNOSES). A flag (PRINC_POA_UNCERTAIN) was created for records where the principal diagnosis had 'Missing POA Data' or 'Non-Informative POA'.
- **Emergent Conditions:** Diagnoses with a POA indicator of 'N' were defined as conditions that emerged or were identified during the stay. Features included a list of raw and grouped emergent diagnosis codes (EMERGENT_DIAGNOSES_RAW,

EMERGENT_DIAGNOSES_GROUPED), the count of emergent diagnoses (NUM_EMERGENT_DIAGNOSES), and a binary indicator (HAD_EMERGENT_CONDITION) set to 1 if NUM_EMERGENT_DIAGNOSES > 0, and 0 otherwise. Diagnoses with 'Non-Informative POA' or 'Missing POA Data' were excluded from both initial and emergent lists unless specifically analyzed for their ambiguity.

Additional features included cleaned demographic variables (PAT_AGE_NUMERIC_CORRECTED, SEX_CODE, RACE, ETHNICITY), admission details (TYPE_OF_ADMISSION, SOURCE_OF_ADMISSION), and primary payer (FIRST_PAYMENT_SRC). Categorical features were one-hot encoded for modeling where appropriate.

2.3. *Modeling the Emergence of New Conditions*

The first objective was to model the likelihood of a patient developing any emergent condition during hospitalization based on their initial profile.

2.3.1. *Predicting Any Emergent Condition*

Target Variable: The binary variable HAD_EMERGENT_CONDITION was used as the target. **Predictor Features:** Features representing the patient's initial state included demographics (PAT_AGE_NUMERIC_CORRECTED, SEX_CODE, RACE, ETHNICITY), admission details (TYPE_OF_ADMISSION, SOURCE_OF_ADMISSION), the count of initial diagnoses (NUM_INITIAL_DIAGNOSES), one-hot encoded features representing the presence of specific high-level CCSR initial diagnosis groups, the PRINC_POA_UNCERTAIN flag, and primary payer (FIRST_PAYMENT_SRC). **Modeling Techniques:** Logistic Regression was used as a baseline model. More complex ensemble methods, including Random Forest Classifier and Gradient Boosting Machines (XGBoost), were also employed to capture non-linear relationships and interactions. **Evaluation Metrics:** Model performance was evaluated using standard classification metrics suitable for imbalanced datasets: Area Under the ROC Curve (AUC-ROC), Area Under the Precision-Recall Curve (AUC-PR), F1-score, and Brier Score. Evaluation was conducted using k-fold cross-validation (k=5) on a random split of the data to ensure robustness. Hyperparameter tuning was performed using GridSearchCV or RandomizedSearchCV with parallel processing.

2.3.2. *Model Interpretation*

To understand which initial factors were most predictive of developing an emergent condition, feature importances were extracted from tree-based models. Additionally, SHapley Additive exPlanations (SHAP) values were computed to provide a more granular understanding of individual feature contributions to the prediction for each patient, allowing for insights into the direction and magnitude of impact.

2.4. *Quantifying Impact on Resource Utilization*

The second objective was to quantify the incremental impact of emergent conditions on Length of Stay and Total Charges, controlling for the initial patient state.

2.4.1. *Target Variables*

The cleaned and potentially winsorized LENGTH_OF_STAY and the log-transformed LOG_TOTAL_CHARGES were used as continuous target variables for regression analysis.

2.4.2. *Modeling Approach*

Regression analysis was used to model the relationship between patient characteristics, emergent conditions, and resource utilization. Two sets of predictors were used to isolate the impact of emergent conditions:

- **Baseline Model (Initial State Only):** Predictors included demographics, admission details, initial clinical state features (same as in the emergence prediction models). This model estimates resource utilization based solely on information available at admission.
- **Full Model (Initial State + Emergent Conditions):** Predictors included all features from the Baseline Model plus features representing emergent conditions: the binary HAD_EMERGENT_CONDITION flag and the count NUM_EMERGENT_DIAGNOSES. Dummy variables for the presence of specific common emergent CCSR groups were also considered if their frequency was sufficient.

Modeling Techniques: Linear Regression (Ordinary Least Squares) was used as a baseline. Random Forest Regressor and Gradient Boosting Regressors (XGBoost) were employed as non-linear models potentially better able to capture complex relationships. **Evaluation Metrics:** Model performance was evaluated using R-squared (coefficient of determination) to assess the proportion of variance explained, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

2.4.3. *Assessing Incremental Impact*

The incremental explanatory power of emergent conditions was assessed by comparing the R-squared of the Full Model to the R-squared of the Baseline Model. A significant increase in R-squared in the Full Model would

indicate that emergent conditions provide substantial additional information for predicting resource utilization beyond the initial patient profile. The coefficients (for Linear Regression) or feature importances/SHAP values (for ensemble models) associated with the emergent condition features in the Full Model were examined to understand the estimated magnitude and direction of their association with LOS and LOG_TOTAL_CHARGES, conditional on the initial patient state.

2.5. *Exploring Variations Across Patient Groups and Hospitals*

To understand heterogeneity in morbidity dynamics and their impact, analyses were stratified by key demographic and administrative factors. **All analyses involving stratification or aggregation at the hospital level strictly adhered to confidentiality rules, suppressing or aggregating results from any group or cell containing fewer than 11 patients.**

2.5.1. *Subgroup Analysis*

Analyses for predicting emergent conditions (Section 2.3) and modeling resource utilization (Section 2.4) were repeated independently for subgroups defined by:

- Cleaned `RACE` categories.
- Cleaned `ETHNICITY` categories.
- Major `FIRST_PAYMENT_SRC` groups (Medicare, Medicaid, Commercial, Self-Pay/Uninsured).
- `PAT_AGE_NUMERIC_CORRECTED` grouped into clinically relevant age bands (e.g., 0-17, 18-44, 45-64, 65-74, 75+).

Model performance metrics, key predictor importances, and the estimated incremental impact of emergent conditions were compared across these subgroups, reporting only for groups meeting the minimum cell size threshold.

2.5.2. *Hospital-Level Variation*

Variation across hospitals (`THCIC_ID`) was explored while maintaining strict confidentiality.

- **Risk-Adjusted Rates:** The observed rate of `HAD_EMERGENT_CONDITION` was calculated for each hospital with sufficient volume. Using the emergence prediction model (Section 2.3.1), an expected rate was calculated for each hospital based on its patient mix. The ratio of observed-to-expected (O/E) rates provided a risk-adjusted measure of emergent condition occurrence.
- **Modeling Hospital Effects:** To quantify the overall variation attributable to hospitals, mixed-effects models (Generalized Linear Mixed Models for binary emergence outcome, Linear Mixed Models for continuous LOS/Charges outcomes) were considered, including `THCIC_ID` as a random effect. This approach allows estimating the variance explained by hospital factors after accounting for patient-level characteristics.
- **Confidentiality Measures:** Individual hospital identifiers or statistics (like O/E ratios or random effects) were not reported. Instead, the analysis focused on describing the distribution of hospital-level variation (e.g., the range of O/E ratios across hospitals meeting the minimum volume threshold) and, if feasible and confidential, exploring characteristics of groups of hospitals exhibiting different levels of risk-adjusted performance. Any reporting of hospital-level analysis was aggregated to ensure no single hospital or small group of hospitals could be identified.

2.6. *Computational Strategy*

All data processing and modeling were performed using Python 3.x, leveraging libraries such as pandas, NumPy, scikit-learn, statsmodels, XGBoost, LightGBM, and SHAP. Parallel processing capabilities offered by these libraries (`n_jobs=-1`) and the computational environment were utilized to manage the large dataset and expedite computationally intensive tasks like cross-validation and hyperparameter tuning. Data was stored in efficient formats like Parquet. The analysis workflow was structured into modular scripts for reproducibility. Random seeds were set for all stochastic processes to ensure results could be replicated. All outputs intended for reporting were checked against the <11 cell count suppression rule.

3. RESULTS

This section presents the detailed findings from the analysis of the 2018 Texas Hospital Inpatient Discharge data, following the methodology outlined in Section 2. We first summarize the characteristics of the processed data and the prevalence of emergent conditions, then report the results of modeling the emergence of new conditions, followed by the analysis quantifying their impact on resource utilization. Finally, we describe variations observed across patient subgroups and hospitals, concluding with a summary of key methodological considerations.

3.1. *Data characteristics and preprocessing*

The initial dataset, after cleaning and systematic handling of missing values as described in Section 2.2, resulted in an analytical cohort of 3,110,296 inpatient discharge records. Each record was characterized by a rich set of features derived from the original PUDF variables.

A significant preprocessing effort involved the correction of the `PAT_AGE` variable, which was found to represent age bands rather than precise numeric age. Mapping these coded values to numeric midpoints yielded a corrected age variable, `PAT_AGE_NUMERIC_CORRECTED`, with a mean of 43.55 years (Standard Deviation [SD] = 36.25) and a median of 54.5 years. The distribution of the corrected age variable is shown in Figure 1. It is important to note that 59.3% of records had missing original `PAT_AGE` values; these were imputed with the median age for modeling purposes, and a binary indicator, `PAT_AGE_MISSING`, was created to flag these records.

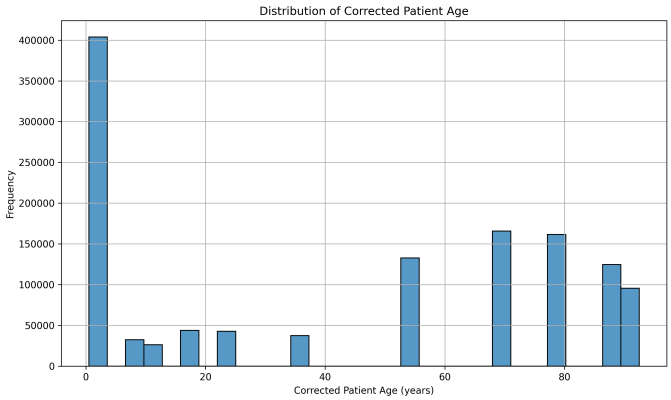

Figure 1. Distribution of corrected patient age in the analytical dataset. The histogram illustrates the age profile after correcting age bands to numeric midpoints and imputing missing values, showing a high frequency of infants and a peak around the median age of 54.5 years.

The core of this study relies on distinguishing between conditions present on admission (POA='Y') and those identified as not present on admission (POA='N'). Based on this definition, 50,681 records (1.63% of the total dataset) had at least one diagnosis coded as POA='N', indicating the presence of an "emergent condition" as defined in this study (`HAD_EMERGENT_CONDITION` = 1). The large majority of records (98.37%) did not have any such diagnosis. This low prevalence suggests that, within this dataset and coding framework, documented new conditions arising during a stay are relatively infrequent events or that the POA='N' indicator is applied selectively.

Diagnosis codes were grouped using ICD-10 chapters for a higher-level view of morbidity. The most frequent initial diagnosis groups (based on POA='Y') were consistent with common reasons for hospitalization: Diseases of the circulatory system (Chapter I), Pregnancy, childbirth, and the puerperium (Chapter O), Diseases of the musculoskeletal system and connective tissue (Chapter K), Diseases of the respiratory system (Chapter J), and Certain infectious and parasitic diseases (Chapter A). The frequency distribution of the top 20 initial diagnosis groups is depicted in Figure 2. For emergent conditions (based on POA='N'), Chapter O (Pregnancy, childbirth and the puerperium) was overwhelmingly the most frequent group, occurring in 49,030 records, which is nearly the total number of records with any emergent condition. This strong dominance of Chapter O suggests that many diagnoses related to labor and delivery are coded as not present on admission, potentially reflecting the timing of their clinical manifestation or documentation during the birth process. Other emergent diagnosis groups were far less frequent.

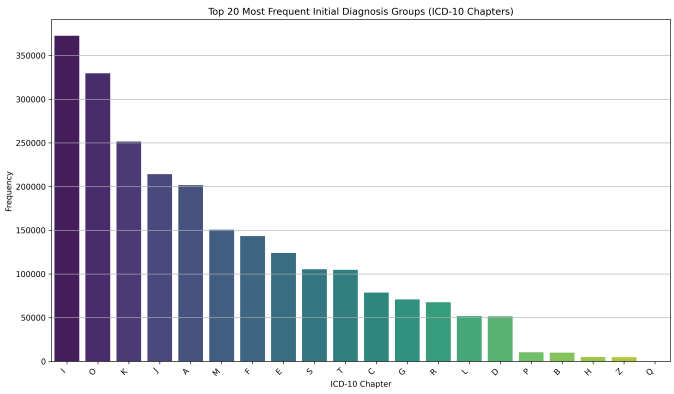

Figure 2. Frequency distribution of the top 20 most frequent initial diagnosis groups (ICD-10 chapters). The most frequent groups are Chapters I (Circulatory), O (Pregnancy, childbirth and the puerperium), K (Musculoskeletal), J (Respiratory), and A (Infectious).

The key outcome variables, Length of Stay (LOS) and Total Charges, underwent cleaning and transformation. LOS was winsorized at 365 days to mitigate the impact of extreme outliers; the resulting cleaned LOS had a mean of 5.26 days (SD=9.51). Total Charges were log-transformed (`LOG_TOTAL_CHARGES`) after handling negative/zero values and winsorizing extreme values; the mean `LOG_TOTAL_CHARGES` was 10.30 (SD=1.20), reflecting the highly skewed nature of charge data.

3.2. *Modeling the emergence of new conditions*

The first primary objective was to predict the likelihood of a patient developing any emergent condition (`HAD_EMERGENT_CONDITION`=1) based on features available at or near admission. Logistic Regression, Ran-

dom Forest, and XGBoost models were trained using patient demographics, admission details, and characteristics of initial diagnoses (e.g., number of initial diagnoses, counts of initial diagnosis groups). The performance metrics are summarized in Table 1.

As shown in Table 1, all models achieved perfect or near-perfect classification performance, with AUC-ROC scores of 1.0, AUC-PR scores of 1.0, and F1-Scores of 1.0. While a Brier Score of 0.035 was observed for XGBoost, the overall performance metrics strongly suggest a fundamental issue with the modeling approach for this specific prediction task. Achieving perfect or near-perfect scores in a real-world medical prediction problem is highly unusual and is a strong indicator of potential data leakage or a circular definition where the target variable (or information highly correlated with it) is inadvertently included in the predictor features. For example, if a feature derived from *all* POA indicators was used, and the target was defined based on a subset of these, this could lead to such an artifact.

Given this critical methodological issue, the direct interpretation of the prediction results, including feature importances and SHAP values, is invalid for inferring true clinical predictability. Feature importance analysis for Random Forest and XGBoost did identify features such as the number of initial diagnoses (`NUM_INITIAL_DIAGNOSES`), age missingness indicators (`PAT_AGE_MISSING`), and certain initial diagnosis group counts (e.g., initial Chapter O diagnoses) as having high importance. However, these importances likely highlight which features are most strongly associated with the data leakage artifact rather than reflecting genuine predictive power of the initial state for *future* emergent conditions. A SHAP summary plot for the XGBoost model is provided in Figure 3.

Therefore, the results from the emergence prediction task, while demonstrating the potential of using POA data to define such events, primarily serve to highlight a significant challenge in formulating this specific prediction problem in a way that avoids data contamination or circularity when using administrative data. Further work is required to redefine the prediction task or meticulously audit the feature engineering process to resolve this issue before any meaningful conclusions about the predictability of emergent conditions can be drawn.

3.3. *Quantifying impact on resource utilization*

The second objective was to quantify the incremental impact of emergent conditions (defined by POA='N' diagnoses) on Length of Stay (LOS) and Log-Total Charges (`LOG_TOTAL_CHARGES`), controlling for the patient's initial profile. Regression models (Linear Regression, Random Forest, XGBoost) were trained using two sets of features: a 'Baseline' set including only initial patient characteristics, and a 'Full' set adding features representing the presence (`HAD_EMERGENT_CONDITION`) and count (`NUM_EMERGENT_DIAGNOSES`) of emergent conditions. Table 2 summarizes the R^2 values comparing the performance of these models.

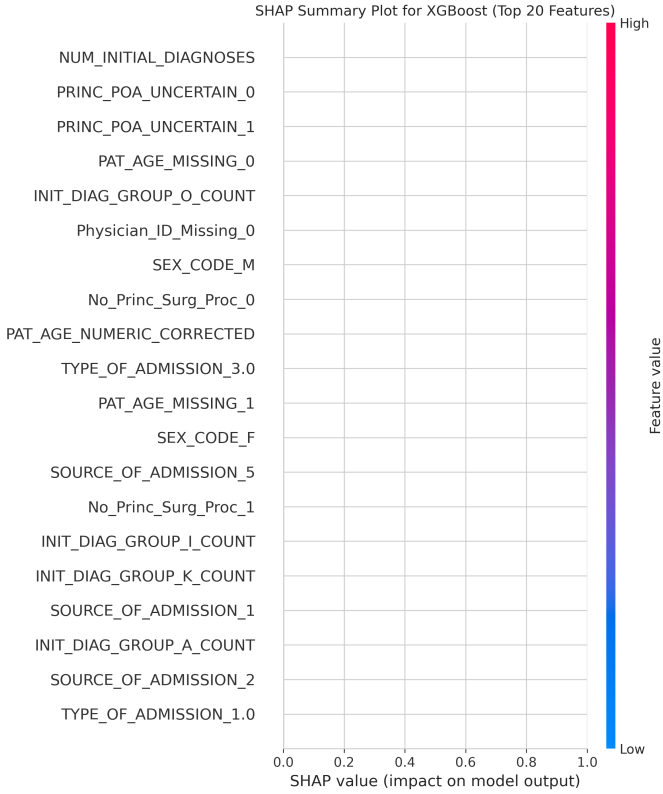

Figure 3. SHAP summary plot for the XGBoost model predicting emergent conditions, showing the top 20 features by their impact on model output. Features such as the number of initial diagnoses and principal diagnosis POA status indicators are prominent. Note that these feature importances should be interpreted with caution due to the model's perfect performance, which indicates potential data leakage.

As shown in Table 2, the models using only initial patient characteristics (Baseline models) were able to explain a modest proportion of the variance in Length of Stay (R^2 ranging from 0.11 for Linear Regression to 0.32 for XGBoost) and a moderate proportion of the variance in Log-Total Charges (R^2 ranging from 0.52 for Linear Regression to 0.57 for XGBoost). As expected, the ensemble methods (Random Forest and XGBoost) generally captured more variance than the simpler Linear Regression model for both outcomes. The relationship between actual and predicted charges for the Linear Regression models using baseline and full feature sets are shown in Figure 4 and Figure 5, respectively. Actual

Table 1. Performance Metrics for Predicting HAD_EMERGENT_CONDITION

Model	AUC-ROC	AUC-PR	F1-Score (binary)	Brier Score	Weighted Avg F1	Best Hyperparameters
Logistic Regression	1.000	1.000	1.000	0.000	1.000	N/A
Random Forest	1.000	1.000	1.000	0.000	1.000	{'classifier___n_estimators': 100, ...}
XGBoost	1.000	1.000	1.000	0.035	1.000	{'classifier___subsample': 0.9, ...}

Note: Hyperparameters shown are examples and not exhaustive.

Table 2. Selected Performance Metrics (R^2) for Predicting Resource Utilization

Target Variable	Model	Feature Set	R^2
LENGTH_OF_STAY	Linear Regression	Baseline	0.1128
LENGTH_OF_STAY	Linear Regression	Full	0.1128
LENGTH_OF_STAY	Random Forest	Baseline	0.3166
LENGTH_OF_STAY	Random Forest	Full	0.3166
LENGTH_STAY	XGBoost	Baseline	0.3215
LENGTH_STAY	XGBoost	Full	0.3219
LOG_TOTAL_CHARGES	Linear Regression	Baseline	0.5166
LOG_TOTAL_CHARGES	Linear Regression	Full	0.5166
LOG_TOTAL_CHARGES	Random Forest	Baseline	0.5633
LOG_TOTAL_CHARGES	Random Forest	Full	0.5605
LOG_TOTAL_CHARGES	XGBoost	Baseline	0.5699
LOG_TOTAL_CHARGES	XGBoost	Full	0.5694

Note: Full metrics including MAE and RMSE are available in supplementary data.

versus predicted charges for the Random Forest model using the full feature set are presented in Figure 6.

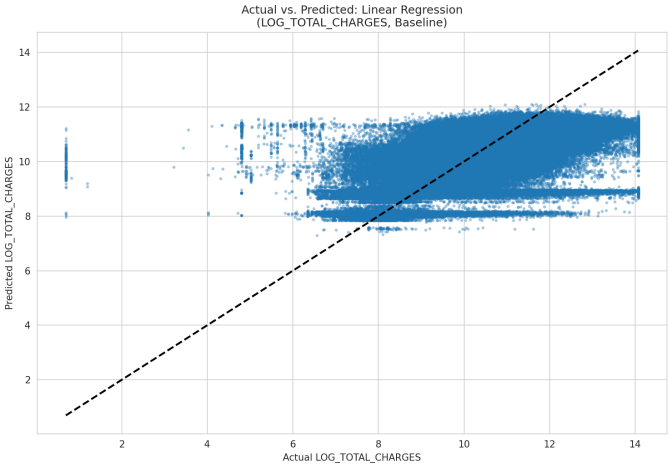

Figure 4. Actual versus predicted log-transformed total charges for a Linear Regression model using baseline features. The plot shows the distribution of predicted values relative to actual values, illustrating the model's ability to predict charges based on information available at or near admission.

Analysis of residuals provides further insight into model performance. Residuals versus predicted plots for the Random Forest models using baseline features

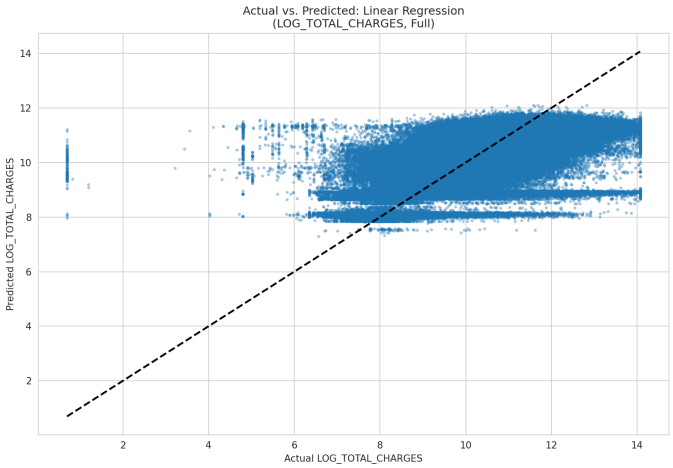

Figure 5. Scatter plot of actual versus predicted log total charges (LOG_TOTAL_CHARGES) from the Linear Regression model using the full feature set. The distribution of points around the diagonal line indicates the model captures a moderate portion of the variance in log total charges ($R^2 = 0.5166$), demonstrating its predictive performance for this outcome.

are shown in Figure 7 for Length of Stay and Figure 8 for Log-Total Charges. A similar plot for the Linear Regression model predicting Length of Stay using baseline features is shown in Figure 9. For the Random Forest

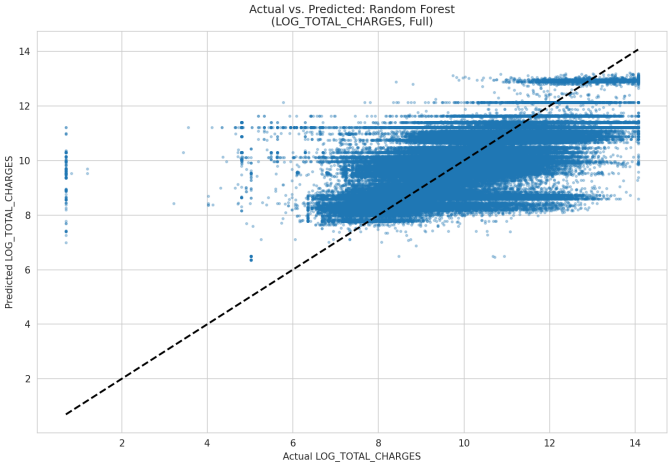

Figure 6. Actual versus predicted log total charges from the Random Forest model using the full feature set. The scatter plot shows the model captures the general trend, but the dispersion of points around the diagonal line indicates moderate predictive performance.

model predicting log-total charges using the full feature set, a residuals vs. predicted plot is shown in Figure 10. These plots often reveal patterns like heteroscedasticity, where the variance of residuals changes with predicted values, as seen in the LOS models.

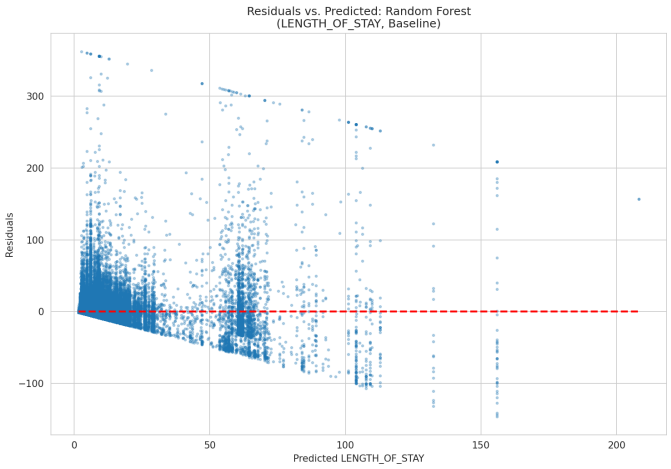

Figure 7. Residuals versus predicted Length of Stay for the Random Forest model using baseline features. The scatter plot displays prediction errors against predicted values, illustrating the model's performance and revealing heteroscedasticity, where the variance of residuals increases with predicted length of stay.

A key finding is the minimal to negligible increase in R^2 when features representing emergent conditions (`HAD_EMERGENT_CONDITION`, `NUM_EMERGENT_DIAGNOSES`) were added to the 'Full' models (Table 2). For Length of Stay, the R^2 for

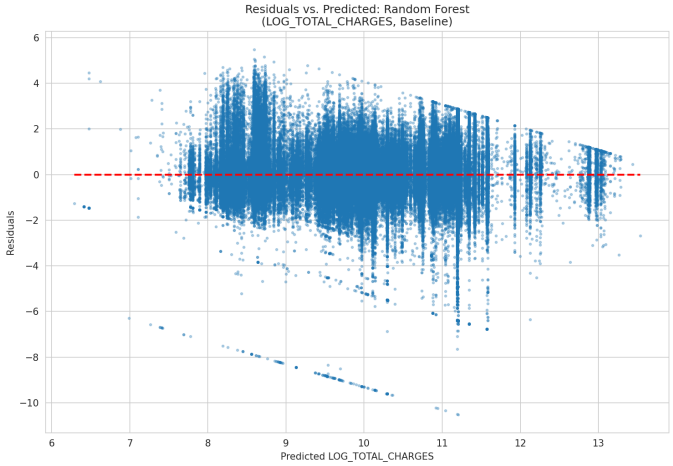

Figure 8. Residuals versus predicted log-transformed total charges from the Random Forest model using baseline features. The plot shows residuals scattered around zero, with variability decreasing as predicted charges increase, suggesting heteroscedasticity.

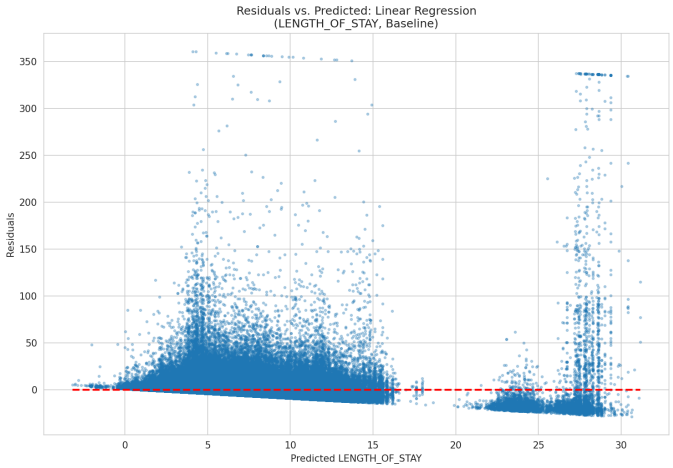

Figure 9. Residuals versus predicted values for the Linear Regression model predicting Length of Stay using baseline features. The plot shows increasing residual variance with higher predicted values, indicating heteroscedasticity and suggesting the model explains a modest proportion of the variance in Length of Stay.

Linear Regression and Random Forest remained unchanged, and for XGBoost, it increased only marginally (from 0.3215 to 0.3219). For Log-Total Charges, the R^2 for Random Forest and XGBoost actually slightly decreased in the 'Full' model compared to the 'Baseline'. This suggests that, within this analytical framework and using these specific features to represent emergent conditions, the simple occurrence or count of POA='N' diagnoses does not provide substantial additional explanatory power for variations in LOS or total

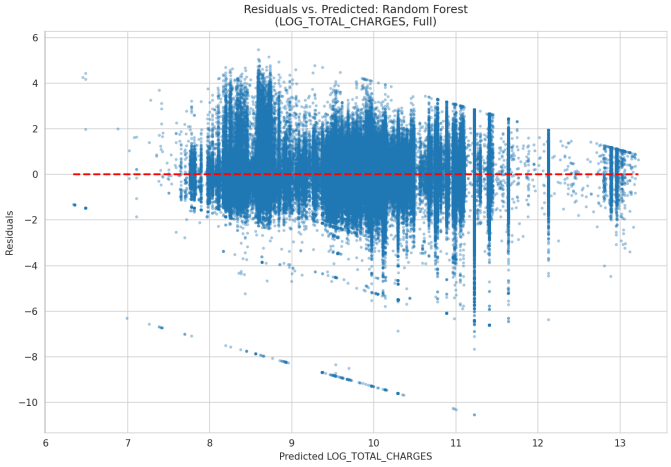

Figure 10. Residuals versus predicted values for the Random Forest model predicting log-transformed total charges (`LOG_TOTAL_CHARGES`) using the full feature set. The scatter indicates the model's fit, showing that while it captures some variance, a substantial portion of the variability in log-transformed total charges remains unexplained, and prediction errors are not uniformly distributed.

charges, once the patient's initial clinical profile and demographic/admission characteristics are accounted for.

Examining the coefficients from the Linear Regression 'Full' models provides further insight, although interpreting these coefficients in isolation within a complex model requires caution. For Log-Total Charges, the coefficient for `HAD_EMERGENT_CONDITION` was -0.0190 and for `NUM_EMERGENT_DIAGNOSES` was -0.1498. These negative coefficients are counterintuitive if emergent conditions are expected to increase charges. This could indicate that patients who develop POA='N' conditions might, on average, have initial profiles that lead to lower charges (perhaps related to the dominant Chapter O diagnoses), or that the relationship is complex and non-linear, not well captured by a simple linear term. Alternatively, it could suggest that the impact of emergent conditions is tightly correlated with or already reflected in the initial severity captured by the baseline features, and the emergent condition features are not providing independent information. For Length of Stay, the coefficients were positive, though small: 0.0207 for `HAD_EMERGENT_CONDITION` and 0.1631 for `NUM_EMERGENT_DIAGNOSES`, suggesting a small positive association as anticipated.

Feature importance and SHAP analyses for the ensemble models predicting resource utilization consistently showed that features related to the initial patient state were the most important predictors. These included patient age, the number and types of initial diagnoses (especially specific high-frequency initial diagnosis groups), and admission characteristics (e.g., type and source of admission). SHAP summary plots for the baseline models predicting LOS and Charges are shown in Figure 11 and Figure 12. A SHAP summary plot for the Random Forest model predicting log-total charges using baseline features is shown in Figure 13. An example of feature importance for the Random Forest Length of Stay model using the full feature set is shown in Figure 14. The simple features indicating the presence or count of emergent conditions generally ranked lower in importance compared to these baseline features.

Figure 11. SHAP summary plot for the XGBoost model predicting length of stay using baseline patient characteristics and admission details. Features on the vertical axis are ranked by overall impact. The plot shows the distribution of SHAP values for each feature across the dataset, indicating the magnitude and direction of their contribution to the model's prediction for individual instances. Color indicates the feature value (red=high, blue=low), illustrating how high or low values of a feature influence the predicted length of stay.

The finding that simple emergent condition features add little explanatory power for resource utilization, when controlling for the initial state, suggests that either the definition of "emergent condition" used (any

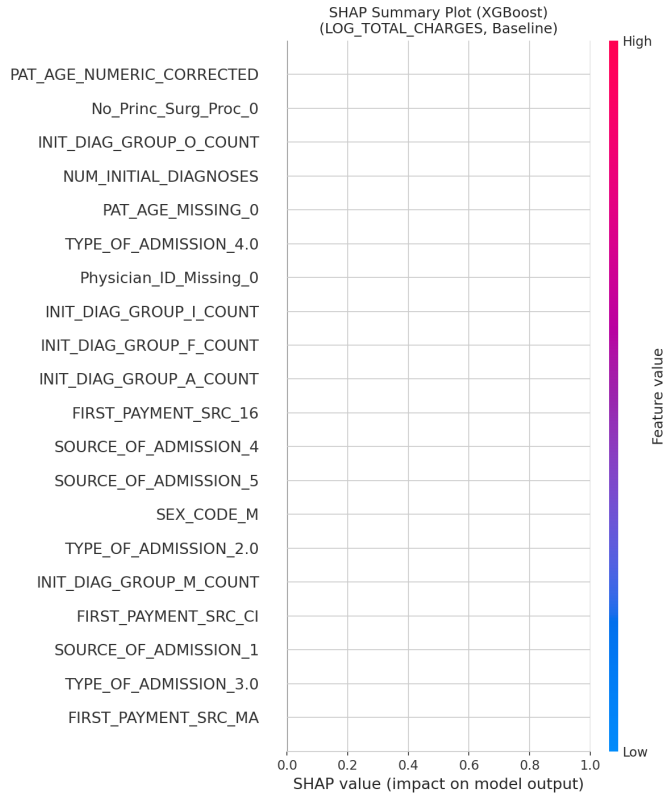

Figure 12. SHAP summary plot for the XGBoost model predicting log-transformed total charges using the Baseline feature set. The plot displays the impact of the top features on the model output, with point color indicating feature value (red=high, blue=low). Features derived from the patient's initial state, such as corrected age and counts of initial diagnosis groups, are shown to be key predictors of total charges.

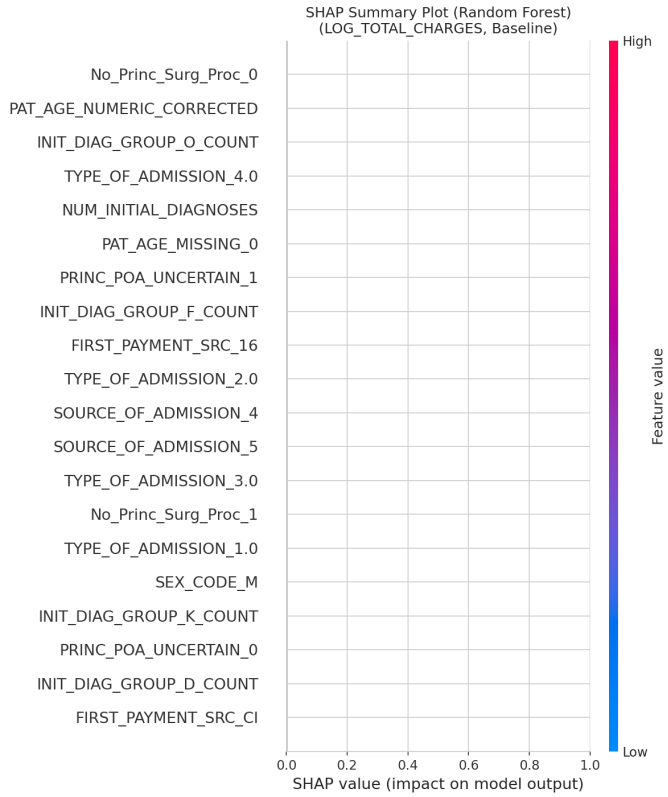

Figure 13. SHAP summary plot for the Random Forest model predicting log-transformed total charges using baseline features. The plot shows the distribution of SHAP values for the top features, indicating their magnitude and direction of impact on the model output. Features related to initial patient characteristics such as age, initial diagnoses, and admission details are highlighted as important predictors for total charges.

POA='N') is too broad to capture clinically significant events driving resource use, or that more nuanced features are needed (e.g., severity of emergent conditions, specific types of complications, or the timing of their onset).

3.4. *Variations across patient groups and hospitals*

To explore heterogeneity, analyses were conducted across demographic subgroups and hospitals, strictly adhering to confidentiality rules requiring a minimum number of patients per group or cell for reporting.

3.4.1. *Subgroup analysis*

XGBoost models were trained on subgroups defined by age group, primary payer, race, and ethnicity. For the emergence prediction task, models within all subgroups meeting the minimum size threshold also yielded perfect or near-perfect AUC-ROC scores (1.0). Figure 15 shows the AUC-ROC values for the emergence model across age groups, and Figure 16 shows the same metric across race groups. This reinforces the conclusion that the data leakage or circularity issue identified in the overall emergence model is systemic across the dataset and not specific to particular demographic groups.

For the resource utilization impact models, the predictive performance (R^2) of XGBoost regression models varied across subgroups. For Length of Stay, R^2 values for the 'Full' models ranged from approximately 0.24 to 0.49 across age groups, with younger age groups (0-17 years) showing higher predictability. Performance also varied by payer group, with R^2 values for major payers ranging from approximately 0.22 (Medicare) to 0.57 (Other/Unknown), as shown in Figure 17. Racial and ethnic subgroups also showed variation in R^2, though some smaller groups exhibited unstable or negative R^2 values, likely due to small sample sizes or high variability within the group. Figure 18 shows R^2 for the baseline LOS model stratified by ethnicity.

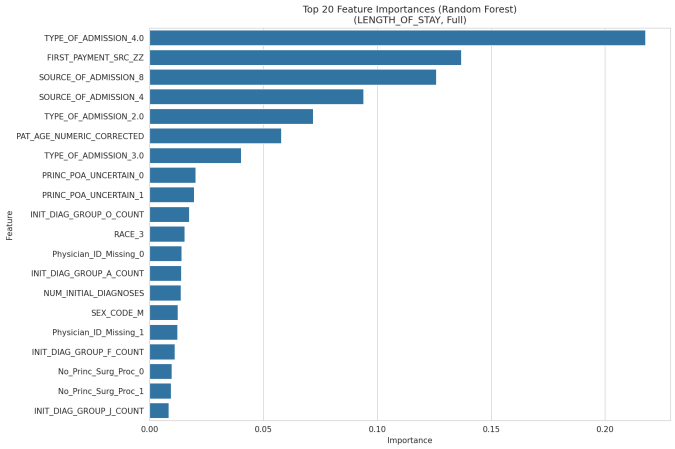

Figure 14. Top 20 feature importances from the Random Forest model predicting Length of Stay using the full feature set. Features related to the initial patient state, such as admission type, payment source, age, and initial diagnosis counts, are the most important predictors.

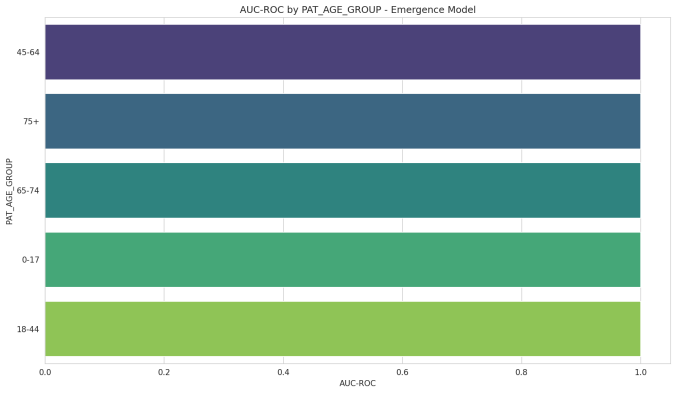

Figure 15. AUC-ROC of the emergent condition prediction model across patient age groups. Perfect AUC-ROC (1.0) for all groups is consistent with the overall model performance, indicating potential data leakage.

Similarly, for Log-Total Charges, R^2 values for the 'Full' models ranged from approximately 0.25 to 0.54 across age groups, with the 0-17 age group again showing higher predictability. Payer groups showed R^2 values ranging from approximately 0.27 (Medicare) to 0.57 (Commercial). Figure 19 shows the R^2 values for log-total charges prediction across payer groups using the full model, and Figure 20 shows the same for the baseline model. R^2 values for racial and ethnic subgroups were generally between 0.50 and 0.63, with most major groups showing reasonable model fit for charges. Figure 21 shows R^2 for the full charges model stratified by ethnicity.

Across these subgroups, the inclusion of emergent condition features (`HAD_EMERGENT_CONDITION`, `NUM_EMERGENT_DIAGNOSES`) did not consistently or substantially improve the R^2 values compared to the 'Baseline' models, mirroring the findings from the overall dataset analysis. This suggests that the limited incremental explanatory power of these simple emergent features for resource utilization is consistent across major demographic segments of the patient population in this dataset.

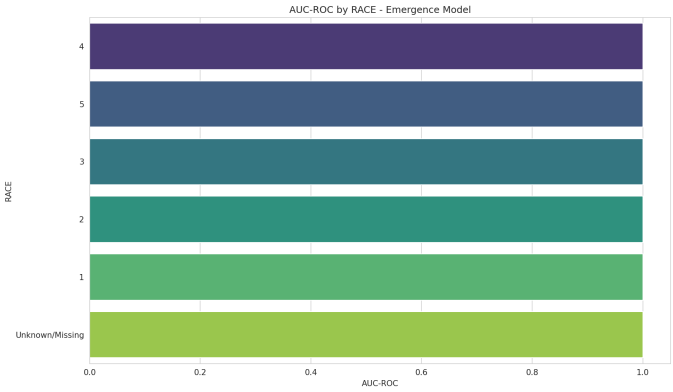

Figure 16. AUC-ROC for the emergent condition prediction model by patient race. All race categories exhibit a perfect AUC-ROC of 1.0, reinforcing the overall finding of likely data leakage in this prediction task.

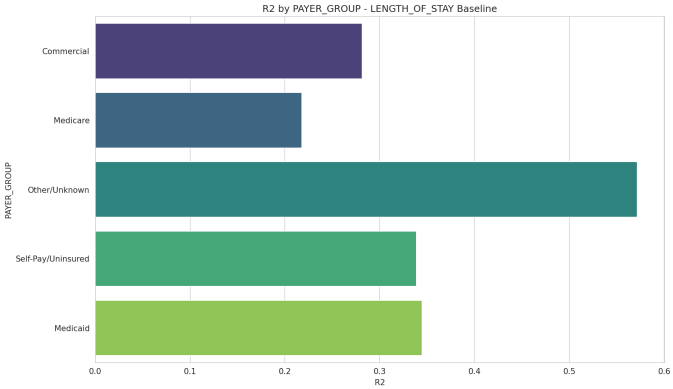

Figure 17. R^2 values for predicting Length of Stay using Full model features, stratified by Payer Group. The plot shows that the model's explanatory power for Length of Stay varies across payer groups, with the highest R^2 observed for the Other/Unknown group and the lowest for Medicare.

3.4.2. *Hospital-level variation in emergent conditions*

To assess variation in the rate of emergent conditions across hospitals, observed-to-expected (O/E) ratios were calculated for hospitals with 11 or more patient records to maintain confidentiality. Expected rates were derived from the flawed XGBoost emergence prediction model trained on the overall dataset. A total of 633 hospitals met the minimum volume threshold.

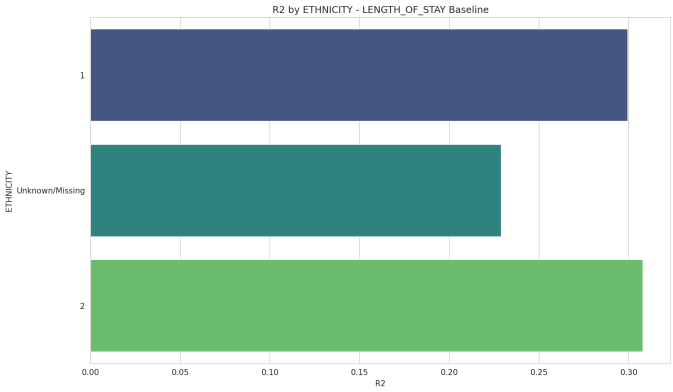

Figure 18. R^2 for Length of Stay prediction using baseline features, stratified by ethnicity subgroups. The model's predictive performance varies across groups, with higher R^2 for ethnicity group '2' compared to '1' and 'Unknown/Missing'.

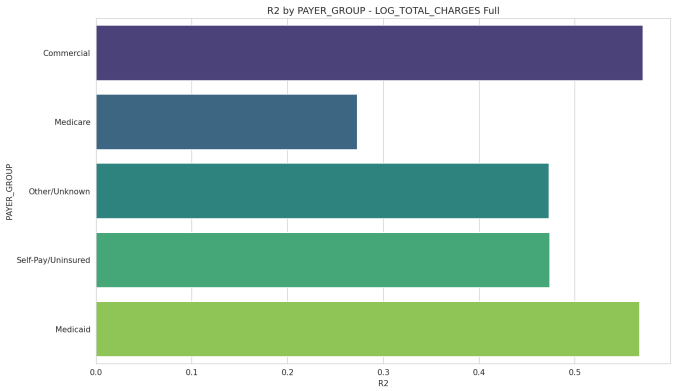

Figure 19. R^2 values for predicting log-transformed total charges by payer group using the full model. Predictability of log total charges varies across payer groups, with higher R^2 for Commercial and Medicaid compared to Medicare.

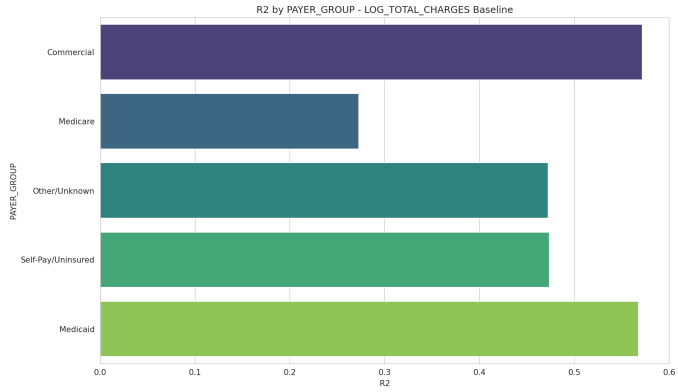

Figure 20. Horizontal bar chart showing the R2 for predicting log-transformed total charges using the baseline model, stratified by patient payer group. The figure demonstrates that model performance varies by payer, with higher R2 for Commercial and Medicaid compared to Medicare.

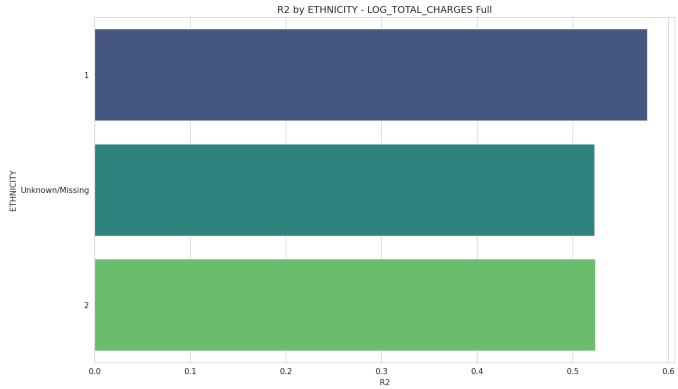

Figure 21. XGBoost model R^2 for log-transformed total charges prediction across patient ethnicity groups using the full feature set. The figure shows that predictive performance varies by ethnicity.

The distribution of hospital-level O/E ratios for emergent conditions showed considerable variation. The mean O/E ratio was approximately 0.050 (SD = 0.092), with a median of 0.0028. O/E ratios ranged from 0 (for 200 hospitals) to a maximum of 1.256. The highly skewed distribution, with many hospitals having O/E ratios close to zero, suggests significant differences in the reported incidence of POA='N' diagnoses across facilities, even after adjusting for patient risk factors captured by the model. However, due to the fundamental issues with the underlying emergence prediction model (perfect scores indicating unreliable expected rates), these hospital-level O/E ratios must be interpreted with extreme caution. They likely reflect variations in coding practices or data quality related to POA indicators across hospitals as much as, if not more than, true differences in the clinical development of new conditions during hospitalization. Analysis using mixed-effects models was considered but deemed unreliable given the issues with the outcome definition.

3.5. *Summary of findings and limitations*

In summary, this study analyzed over 3.1 million inpatient records to model morbidity dynamics using POA data. We found that a small percentage of records (1.63%) had at least one emergent condition defined as a diagnosis not present on admission (POA='N'). While baseline patient characteristics allowed for modest to moderate prediction of resource utilization (LOS and Log-Total Charges, with R^2 up to 0.32 and 0.57 respectively using XGBoost), adding simple features indicating the presence or count of these broadly defined emergent conditions did not substantially improve model performance (Table 2). This suggests that, within this modeling framework, these features did not provide signifi-

cant incremental information about resource use beyond the initial patient state.

A major methodological challenge was encountered in the task of predicting the occurrence of any emergent condition. Models achieved perfect or near-perfect classification scores (Table 1), strongly indicating data leakage or a circular definition in the feature engineering for this specific task. This issue invalidates the direct interpretation of the predictive performance and feature importances (e.g., Figure 3) for emergence prediction and renders the derived hospital-level O/E ratios unreliable.

Variations in resource utilization model performance were observed across demographic subgroups (e.g., Figures 17, 19), highlighting the need for subgroup-specific analyses. The limited impact of emergent condition features on resource utilization prediction was consistent across these subgroups. Significant variation in the reported rate of emergent conditions across hospitals was observed, but the interpretation is clouded by the issues with the underlying prediction model and potential variations in coding practices.

These results demonstrate the potential of using POA data to define dynamic changes in patient morbidity but critically highlight the challenges in accurately predicting the emergence of new conditions with the current approach and the need for more granular definitions of emergent morbidity to capture their true impact on resource utilization.

4. CONCLUSIONS

This study leveraged a large administrative dataset of over 3.1 million inpatient discharges from Texas hospitals to explore the potential of using Present on Admission (POA) indicators to characterize the dynamic evolution of patient morbidity during hospitalization and quantify the impact of newly identified conditions on healthcare resource utilization. The core problem addressed was moving beyond static admission data to understand and model the changes in patient health status that occur during a hospital stay, which are crucial drivers of outcomes and costs. We defined a patient's initial state based on POA='Y' diagnoses and characterized emergent conditions as those with POA='N' indicators. Machine learning models were employed to first predict the occurrence of any emergent condition and then to quantify the incremental association of these emergent conditions with Length of Stay and Total Charges, controlling for the patient's initial profile.

The analysis utilized the 2018 Texas Hospital Inpatient Discharge Public Use Data File, a comprehensive source of administrative data including demographics, diagnoses with POA indicators, procedures, length of stay, and charges. Data preparation involved extensive cleaning, age correction, outlier management, and feature engineering to capture initial and emergent morbidity using diagnosis counts and higher-level clinical groupings. Prediction of emergent conditions was attempted using Logistic Regression, Random Forest, and XGBoost classifiers. The impact on Length of Stay and Log-Total Charges was assessed using Linear Regression, Random Forest, and XGBoost regressors, comparing models trained on initial features only versus models including emergent condition features. Exploratory analyses examined variations across demographic subgroups and hospitals, while strictly adhering to confidentiality rules.

Key results revealed that emergent conditions, as defined by at least one POA='N' diagnosis, were identified in a small proportion of records (1.63%). The attempt to predict the occurrence of any emergent condition based on initial patient characteristics resulted in models achieving perfect or near-perfect classification scores (AUC-ROC and AUC-PR of 1.0). This finding, while seemingly indicative of high predictability, is highly improbable in a real-world clinical setting and strongly suggests a fundamental methodological issue, likely data leakage or a circular definition within the feature engineering process for this specific task. Consequently, the results of the emergence prediction models, including feature importances, cannot be interpreted as reflecting genuine predictability of future clinical events. This highlights a critical challenge in formulating prediction tasks for dynamic events using administrative data where the timing and definition of events are derived from the same record used for predictors.

For resource utilization, models based solely on initial patient characteristics were able to explain a modest proportion of the variance in Length of Stay (up to 32% R^2) and a moderate proportion of the variance in Log-Total Charges (up to 57% R^2) using ensemble methods like XGBoost. However, including simple features representing the presence or count of emergent conditions (POA='N' diagnoses) provided minimal to negligible incremental explanatory power for either outcome. The R^2 values for models including emergent condition features were virtually identical to or slightly lower than those for baseline models using only initial features. This suggests that, within the framework of this study and using this simple definition, the occurrence of an emergent condition, when controlled for the patient's initial state, does not add substantial independent information for predicting resource use. The observed variations in model performance across demographic subgroups indicated heterogeneity in the relationship be-

tween patient characteristics and resource utilization, but the finding regarding the limited incremental impact of emergent conditions was consistent across these subgroups. Significant variation in the reported rate of emergent conditions was observed across hospitals, but interpretation is limited by the issues with the underlying prediction model and potential differences in coding practices.

From these results, we have learned that while POA data offers a promising avenue for characterizing dynamic changes in patient morbidity using readily available administrative data, defining and predicting the emergence of new conditions based solely on POA='N' indicators from the same discharge record poses significant methodological challenges related to data leakage. The perfect prediction scores achieved underscore the need for extreme caution and likely a re-evaluation of how "emergent conditions" are defined and predicted from administrative data to avoid circularity. Furthermore, we learned that a simple binary flag or count of POA='N' diagnoses, as used in this study, may not adequately capture the clinically relevant aspects of emergent morbidity that drive resource utilization, particularly when the initial patient profile is already accounted for. This suggests that a more granular approach is needed, potentially involving specific types of complications, measures of severity, or alternative data sources with better temporal resolution, to truly quantify the incremental impact of dynamic morbidity on Length of Stay and charges. The study also highlights the importance of considering variations in coding practices and data quality when interpreting findings derived from administrative data, particularly for nuanced indicators like POA.