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

Evaluating Attention-Based Learning of Patient Diagnosis Representations with Present On Admission Status for In-Hospital Mortality and Prolonged Length of Stay Prediction

Denario[1]

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

ABSTRACT

Predicting in-hospital outcomes such as mortality and prolonged length of stay using administrative hospital discharge records is crucial for risk stratification and resource management, requiring effective methods to leverage complex clinical information like diagnosis codes and their Present On Admission (POA) status. We developed a novel deep learning approach utilizing a Transformer encoder to learn contextualized patient representations from their set of diagnosis codes, where each diagnosis input token explicitly encodes both the diagnosis identity (truncated ICD-10-CM) and its associated POA status, including a distinct category for missing POA information. This learned patient embedding was then concatenated with other admission-time features including demographics, admission type, and an engineered count of diagnoses present on admission. Using data from the 2018 Texas Hospital Inpatient Discharge Public Use Data File, we trained and evaluated Logistic Regression and Gradient Boosting models on these combined features for predicting in-hospital mortality and prolonged length of stay, comparing performance against baseline models using only non-diagnostic features or simpler, explicit diagnosis encodings. While the attention-based encoder learned representations that captured some predictive signal in a proxy task, final prediction models incorporating these embeddings did not outperform baseline models, particularly those utilizing a simpler encoding of top diagnosis codes alongside other features, for either outcome. The number of diagnoses present on admission was consistently identified as a highly influential predictor across models. These findings suggest that while complex deep learning methods can learn representations from diagnosis-POA sequences, their effectiveness is highly dependent on sufficient training data (limited in this study by data subsampling for the Transformer) and careful integration with other relevant clinical features; simpler feature engineering approaches can provide strong performance baselines.

1. INTRODUCTION

Accurate and timely prediction of critical in-hospital outcomes, such as mortality and prolonged length of stay (PLOS), is a fundamental challenge in healthcare. Such predictions are vital for proactive risk stratification, enabling healthcare providers to identify patients at high risk of adverse events and implement targeted interventions. Furthermore, reliable outcome prediction supports efficient hospital resource management, including staffing, bed allocation, and discharge planning, ultimately contributing to improved patient care and operational efficiency.

Hospital administrative discharge records constitute a rich and widely available source of data for developing predictive models. These records contain a wealth of information captured at or near the time of admission, including patient demographic characteristics, details about the admission itself (e.g., type and source), and, crucially, comprehensive lists of diagnosis codes assigned during the patient's hospital stay. Diagnosis codes, standardized through systems like the International Classification of Diseases (ICD), provide a structured summary of a patient's medical conditions and are central to understanding their clinical profile. However, effectively utilizing the full complexity of diagnosis data for prediction presents significant challenges. Patients often have multiple diagnoses, forming a variable-length set of codes where the interactions and relative importance of individual conditions can be highly complex. The sheer scale of possible diagnosis codes (tens of thousands) results in high-dimensional and sparse data representations that are difficult for traditional models to handle directly.

A particularly challenging, yet potentially informative, aspect of diagnosis data is the Present On Admission (POA) status indicator. POA status specifies whether a given diagnosis was present at the time the patient was admitted to the hospital or if it developed subsequently during the hospitalization. This distinction

carries significant clinical weight, as conditions present on admission are typically the primary reasons for hospitalization and major drivers of resource utilization, while hospital-acquired conditions may signal complications or quality of care issues. Despite its importance, POA status is frequently recorded inconsistently or left blank in administrative data, introducing an additional layer of complexity and uncertainty that must be effectively modeled to avoid misinterpreting the data. Traditional methods often simplify diagnosis information by focusing solely on the principal diagnosis, counting the total number of diagnoses, or using simple "bag-of-codes" approaches, which may fail to capture the nuanced relationships between multiple diagnoses and their associated POA contexts.

To address these challenges and better leverage the detailed information available in diagnosis codes and their POA statuses, we propose and evaluate a novel deep learning approach. Our method centers on learning a contextualized, fixed-size vector representation for each patient directly from their set of diagnosis codes and associated POA indicators. We employ an attention-based architecture, specifically a Transformer encoder, which is well-suited for processing sets of inputs by allowing the model to dynamically weigh the importance and relationships among different diagnosis-POA pairs within a patient's record. A key innovation is the explicit encoding of each diagnosis input token to include both the diagnosis code identity (using truncated ICD-10-CM codes) and its corresponding POA status. Critically, we treat missing POA information not as a simple data omission, but as a distinct category with its own representation, acknowledging the potential informational value (or lack thereof) associated with undocumented POA status.

The learned patient embedding produced by the Transformer encoder is designed to capture the complex interplay of multiple diagnoses and their POA context. This powerful, data-driven representation is then combined with other standard admission-time features, such as demographics, admission type, and an engineered feature representing the count of diagnoses explicitly marked as present on admission. This comprehensive set of features serves as the input for training state-of-the-art machine learning classifiers, specifically Logistic Regression and Gradient Boosting models, to predict in-hospital mortality and prolonged length of stay.

We evaluate the effectiveness of this attention-based diagnosis representation learning approach using a large, real-world dataset derived from the 2018 Texas Hospital Inpatient Discharge Public Use Data File. We compare the performance of predictive models incorporating our learned diagnosis-POA representations against baseline models that utilize only non-diagnostic features or employ simpler, more traditional explicit encodings of diagnosis codes. Through this rigorous evaluation, we aim to determine whether the proposed complex representation learning method offers significant predictive advantages for critical in-hospital outcomes and to identify the most influential predictors within this data context.

2. METHODS

2.1. *Data Source and Study Population*

This study utilized data from the 2018 release of the Texas Hospital Inpatient Discharge Public Use Data File (PUDF). This administrative dataset contains records for inpatient hospital stays in Texas, including patient demographics, admission and discharge details, diagnoses coded using the International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM), procedure codes, and associated Present On Admission (POA) indicators. The raw dataset comprised 3,110,296 patient discharge records. For this analysis, we included all records present in the dataset after initial loading and removal of a single entirely null column identified during exploratory data analysis.

2.2. *Outcome Variables*

Two primary in-hospital outcomes were defined and predicted: In-Hospital Mortality and Prolonged Length of Stay (PLOS).

2.2.1. *In-Hospital Mortality*

In-hospital mortality was defined as a binary variable, `MORTALITY`. A patient record was assigned a value of 1 if the Patient Discharge Disposition (`PAT_STATUS`) was recorded as '20' (Expired), indicating the patient died during the hospitalization. Otherwise, `MORTALITY` was set to 0. This yielded a mortality rate of 1.74% in the dataset, reflecting the inherent class imbalance of this outcome.

2.2.2. *Prolonged Length of Stay*

Prolonged Length of Stay (`PLOS`) was also defined as a binary variable. Based on exploratory data analysis which identified the 75th percentile of `LENGTH_OF_STAY` as 6 days, `PLOS` was set to 1 if `LENGTH_OF_STAY` was greater than 6 days, and 0 otherwise. Records with missing `LENGTH_OF_STAY` (a rate of 0.003%) were excluded from the PLOS prediction task. Using a percentile threshold for dichotomization mitigates the influence of extreme outliers in the original `LENGTH_OF_STAY` variable on the binary outcome definition.

2.3. *Feature Engineering*

Features were engineered exclusively from variables known at the time of admission to ensure that predictions were based on available information at the start of the hospital stay. These features were broadly categorized into non-diagnosis features and diagnosis-related features, with a novel approach developed for the latter.

2.3.1. *Non-Diagnosis Features*

Admission-time features included demographics (`PAT_AGE`, `SEX_CODE`, `RACE`, `ETHNICITY`), geographic information (`PAT_COUNTY`, `PAT_ZIP`), and admission details (`TYPE_OF_ADMISSION`, `SOURCE_OF_ADMISSION`).

Categorical variables were processed as follows:

- `PAT_AGE` was mapped to predefined age groups (`PAT_AGE_mapped`) based on the dataset's specifications.
- `SEX_CODE` was mapped to 'F', 'M', and 'U' (Unknown), treating 'U' as a distinct category due to its non-random nature.
- `RACE` and `ETHNICITY` were mapped to standardized categories, with rare or missing values grouped into 'Other' or 'Unknown' categories.
- `TYPE_OF_ADMISSION` and `SOURCE_OF_ADMISSION` had very low missingness and were imputed with the mode.
- Missing values for `PAT_ZIP` (8.62%) and `PAT_COUNTY` were assigned specific 'MISSING' categories.

These categorical features (`PAT_AGE_mapped`, `SEX_CODE_mapped`, `RACE_mapped`, `ETHNICITY_mapped`, `TYPE_OF_ADMISSION`) were one-hot encoded. Due to their high cardinality, `PAT_COUNTY` and `PAT_ZIP` were handled differently. `PAT_ZIP` codes were first truncated to their first three digits (ZIP3). Both `PAT_COUNTY` and ZIP3 were then encoded using target encoding, calculating encoding values based on the outcome prevalence within each category on the training data with smoothing applied to mitigate the impact of low-frequency categories.

A numerical feature, `NUM_POA_Y_DIAGNOSES`, was engineered by counting the total number of diagnoses (Principal and Other Diagnoses 1-24) for each patient that were explicitly marked with a POA status of 'Y'. This feature was standardized using z-score normalization.

2.3.2. *Diagnosis and Present On Admission (POA) Processing*

The comprehensive list of diagnosis codes (`PRINC_DIAG_CODE` and `OTH_DIAG_CODE_1` through `OTH_DIAG_CODE_24`) and their associated POA indicators (`POA_PRINC_DIAG_CODE`, `POA_OTH_DIAG_CODE_1`, etc.) formed the basis for learning patient diagnosis representations. All ICD-10-CM codes were truncated to their first three characters to reduce dimensionality while retaining clinically relevant categorical information. Null or empty diagnosis code fields were treated as absence of a diagnosis at that position.

POA statuses were standardized into six distinct categories based on the raw data values: 'Y' (Yes), 'N' (No), 'U' (Unknown), 'W' (Cannot be determined), '1' (Exempt), and critically, a specific category `POA_M_MISSING` was assigned to instances where a diagnosis code was present but the corresponding POA field was blank or null. This approach explicitly models the informational distinction of missing POA data, which was prevalent in the dataset. Any other invalid POA codes were mapped to `POA_M_MISSING`.

For each patient, a set of diagnosis-POA pairs was constructed. This set included the pair (`truncated_PRINC_DIAG_CODE`, `processed_POA_PRINC_DIAG_CODE`) and, for each j from 1 to 24, if `OTH_DIAG_CODE_j` was not null, the pair (`truncated_OTH_DIAG_CODE_j`, `processed_POA_OTH_DIAG_CODE_j`). Pairs with null diagnosis codes were excluded.

2.4. *Diagnosis Set Representation Learning via Attention Mechanism*

A core component of this study was the development of a deep learning model, specifically a Transformer encoder, to learn a fixed-size vector representation for each patient from their structured set of diagnosis-POA pairs. This aimed to capture the complex interactions and context among multiple diagnoses and their POA statuses, addressing limitations of simpler bag-of-codes approaches discussed in the introduction.

2.4.1. *Input Tokenization and Embedding*

A vocabulary of unique 3-character diagnosis codes was created from the training data, with codes occurring below a frequency threshold mapped to an 'UNK_DX' token. Similarly, a vocabulary of the six processed POA status categories was created.

Each diagnosis-POA pair for a patient was treated as an input token. An embedding layer mapped each unique diagnosis code to a dense vector of dimension D_{dx}, and another embedding layer mapped each

unique POA status to a dense vector of dimension D_{poa}. The embedding for a specific diagnosis-POA token was formed by concatenating the corresponding diagnosis code embedding and POA status embedding, resulting in a vector of dimension $D_{dx} + D_{poa}$.

2.4.2. *Transformer Encoder Architecture*

For each patient, the sequence of concatenated diagnosis-POA token embeddings was fed into a Transformer encoder. Input sequences were padded to a maximum length (25 tokens, corresponding to the principal diagnosis and 24 other diagnoses) to ensure uniform input size for batch processing. The Transformer encoder consisted of multiple layers, each comprising a multi-head self-attention mechanism followed by a position-wise feed-forward network, with layer normalization and residual connections applied throughout. The self-attention mechanism allowed the model to weigh the importance and relationships among different diagnosis-POA tokens within a patient's record, effectively learning contextualized representations for each input token. Hyperparameters such as the number of layers, number of attention heads, and embedding/hidden dimensions were tuned during model development.

2.4.3. *Patient Representation Aggregation*

The Transformer encoder outputted a sequence of context-aware embeddings, one for each input diagnosis-POA token. To obtain a single, fixed-size patient-level diagnosis representation vector, these output embeddings were aggregated using mean pooling across the sequence dimension. This resulting vector served as the learned diagnosis-POA embedding for the patient.

2.5. *Predictive Modeling*

The learned patient diagnosis representation vector was concatenated with the vector of engineered non-diagnosis features (one-hot encoded demographics and admission details, target-encoded geographic features, and the standardized `NUM_POA_Y_DIAGNOSES`). This combined feature vector was then used as input to train two types of classification models for predicting In-Hospital Mortality and PLOS:

2.5.1. *Logistic Regression*

A Logistic Regression model was trained as a simple yet interpretable linear classifier on the combined feature set. L1 or L2 regularization was applied to prevent overfitting.

2.5.2. *Gradient Boosting Machine*

A Gradient Boosting Machine (specifically XGBoost or LightGBM) was trained as a powerful non-linear classifier. This model is well-suited for tabular data and capable of capturing complex interactions among features.

These models were trained independently for the Mortality and PLOS prediction tasks.

2.6. *Baseline Models*

To assess the value of the proposed attention-based diagnosis representation, the performance of the models trained on the combined feature set was compared against two sets of baseline models:

2.6.1. *Baseline 1: Non-Diagnosis Features Only*

Logistic Regression and Gradient Boosting models were trained using only the engineered non-diagnosis features (demographics, admission details, geographic features, and `NUM_POA_Y_DIAGNOSES`), completely excluding diagnosis information.

2.6.2. *Baseline 2: Non-Diagnosis Features + Simple Diagnosis Encoding*

Logistic Regression and Gradient Boosting models were trained using the non-diagnosis features combined with a simpler, more traditional encoding of diagnosis information. This included one-hot encoding of the top N (e.g., 200-300) most frequent 3-character Principal Diagnosis codes, and a multi-hot encoded binary matrix indicating the presence of the top N most frequent 3-character diagnosis codes that were explicitly marked as POA='Y'.

2.7. *Model Training and Evaluation*

The full dataset was split into training, validation, and test sets using a 70%/15%/15% split. Stratification was applied based on both the Mortality and PLOS outcomes to ensure similar prevalence in each split. All model development, including vocabulary creation, embedding training, feature engineering calculations (like target encoding), and hyperparameter tuning, was performed exclusively on the training and validation sets. The held-out test set was used only for final, unbiased evaluation of the trained models.

For the In-Hospital Mortality task, given the significant class imbalance, class weighting was applied during model training (e.g., using `class_weight='balanced'` for Logistic Regression or `scale_pos_weight` for Gradient Boosting) to give higher importance to the minority class (mortality).

Hyperparameter tuning for all models (Transformer architecture, Logistic Regression regularization, Gradient Boosting parameters) was performed using cross-validation on the training set, leveraging the avail-

able CPU resources (`n_jobs=-1`) for parallel processing where supported (e.g., by scikit-learn's search utilities).

Model performance was evaluated using standard classification metrics. For the imbalanced Mortality task, the primary evaluation metric was the Area Under the Precision-Recall Curve (AUC-PR), which is more informative than AUC-ROC when the positive class is rare. Secondary metrics for Mortality included AUC-ROC, F1-score, Precision, Recall (Sensitivity), Specificity, and Brier Score. For the PLOS task, the primary evaluation metric was AUC-ROC, with AUC-PR, F1-score, Precision, and Recall as secondary metrics. All reported metrics were calculated on the held-out test set.

2.8. *Model Interpretation*

To gain insights into the factors influencing model predictions, Shapley Additive exPlanations (SHAP) were applied to the best-performing final models (Logistic Regression and Gradient Boosting trained on combined features). SHAP values were used to understand the global importance of features (including the overall contribution of the learned diagnosis representation) and to provide local explanations for individual patient predictions.

3. RESULTS

This section presents the detailed quantitative and qualitative results derived from applying the developed models to the 2018 Texas Hospital Inpatient Discharge Public Use Data File. The primary focus is the comparative performance of the attention-based diagnosis representation learning approach against simpler baselines for predicting in-hospital mortality and prolonged length of stay (PLOS).

3.1. *Data characteristics and preprocessing summary*

The analysis began with 3,110,296 inpatient discharge records. After initial cleaning, 167 features were available. A 1% random subsample of 31,102 records was extracted for training the computationally intensive Transformer encoder, subsequently split into 70% training (21,770), 15% validation (4,666), and 15% test (4,666) sets.

The target variables, In-Hospital Mortality and PLOS, were defined as binary outcomes. Mortality occurred in 1.74% of the full dataset records, highlighting the significant class imbalance addressed during modeling. PLOS, defined as a length of stay greater than 6 days (the 75th percentile), affected approximately 21.1% of admissions. The distribution of the original length of stay is highly right-skewed, as shown in Figure 1, while the distribution of the binary PLOS target variable is illustrated in Figure 2.

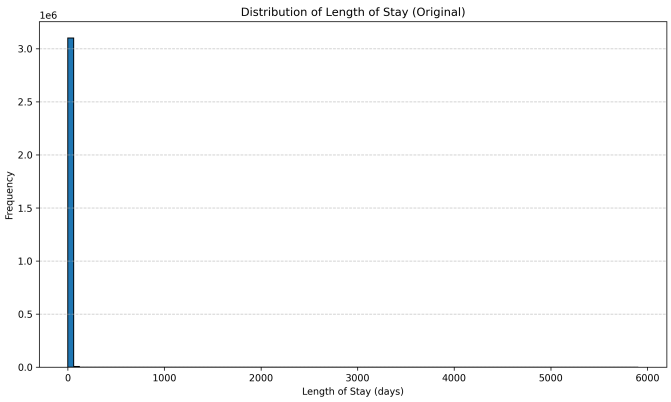

Figure 1. Histogram of original patient length of stay, showing a highly right-skewed distribution with most stays being short.

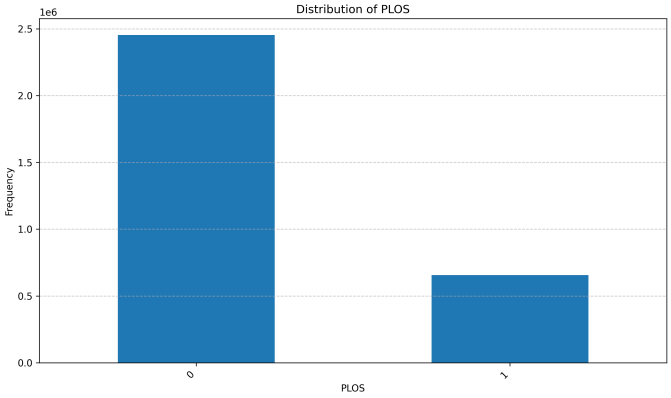

Figure 2. Distribution of the Prolonged Length of Stay (PLOS) target variable. The bar chart shows the frequency count for records classified as not having prolonged stay (PLOS=0) and having prolonged stay (PLOS=1). Approximately 21.1% of records in the dataset are classified as PLOS.

Non-diagnosis features, including demographics, admission details, and geographic information, were processed through mapping, standardization, one-hot encoding, and target encoding as described in the Methods. Key demographic distributions in the processed data are shown in Figure 7 (Age), Figure 6 (Race), and Figure 5 (Ethnicity). Admission characteristics like Type of Admission and Source of Admission are presented in Figure 3 and Figure 4, respectively.

A key engineered feature, the count of diagnoses explicitly marked as Present On Admission ('Y'), denoted as `NUM_POA_Y_DIAGNOSES`, was calculated and standardized. This feature had a mean of 7.36 (SD 5.85) in the full dataset, and its distribution is shown in Figure 8.

Diagnosis codes were truncated to their first three characters. Present On Admission (POA) statuses were standardized into six categories, including a distinct cat-

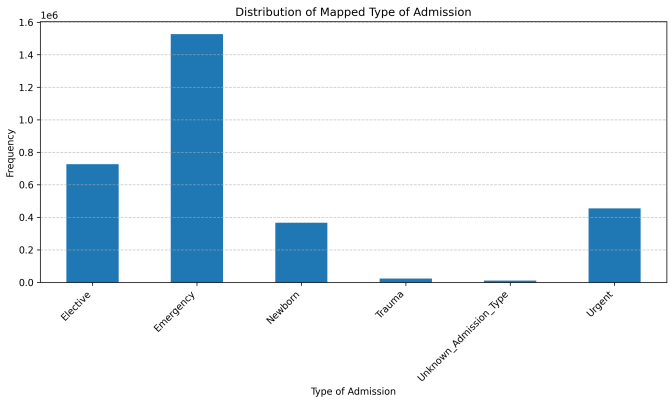

Figure 3. Frequency distribution of mapped Type of Admission categories, indicating that Emergency admissions are the most prevalent, followed by Elective and Urgent admissions.

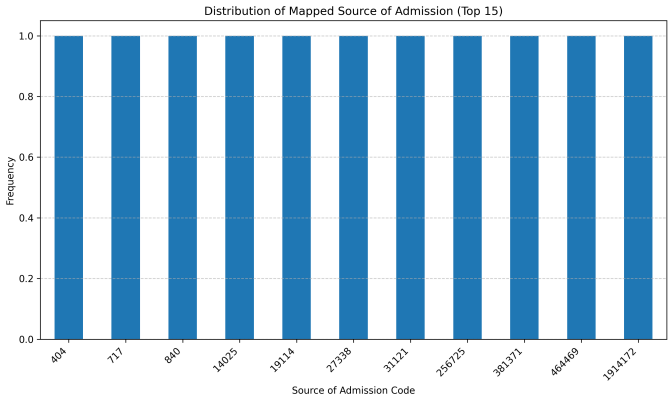

Figure 4. Distribution of the top 15 mapped Source of Admission codes selected during feature processing. The plot shows a frequency of 1.0 for each of these codes.

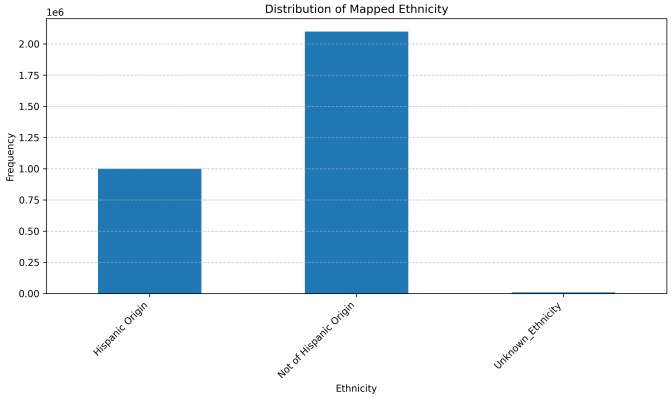

Figure 5. Distribution of mapped ethnicity categories. The histogram shows the frequency of patients categorized as 'Hispanic Origin', 'Not of Hispanic Origin', and 'Unknown_Ethnicity', illustrating the demographic composition of the dataset after preprocessing.

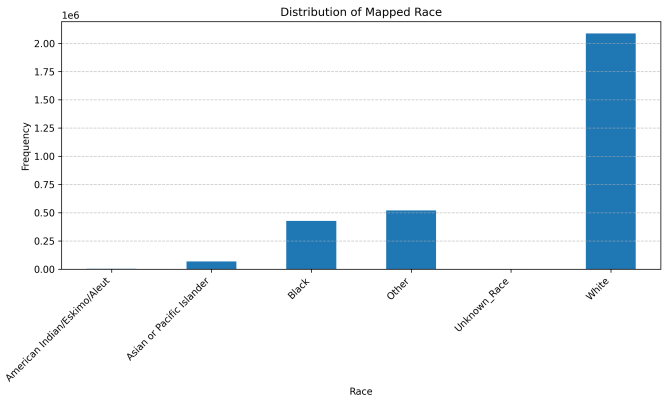

Figure 6. Frequency distribution of mapped patient race categories in the dataset. The plot shows the prevalence of different race categories after preprocessing, with 'White' being the most frequent.

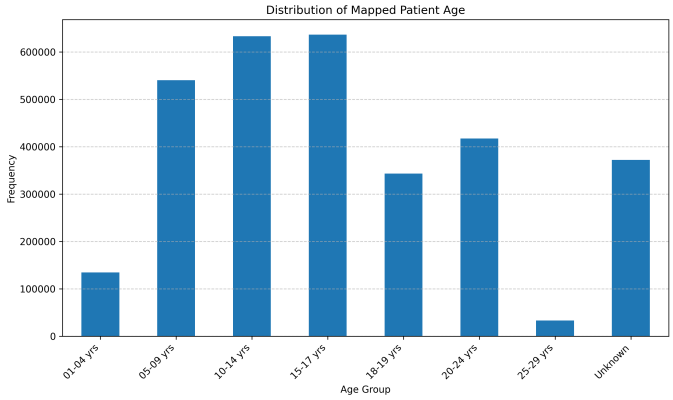

Figure 7. Distribution of patient age groups. The bar chart shows the frequency of patients within each mapped age category, illustrating the prevalence of different age ranges and the number of patients with unknown age.

egory for missing information (`POA_M_MISSING`), which was notably prevalent in the dataset. The distribution of processed POA statuses for the principal diagnosis is shown in Figure 9.

3.2. *Diagnosis_POA sequence construction and vocabulary*

For the attention-based model, patient inputs were constructed as sequences of (truncated diagnosis code, processed POA status) pairs, up to a maximum length of 25. A vocabulary of 1,560 unique 3-character diagnosis codes (including 'UNK_DX' and 'PAD_DX') was derived, covering the majority of observed codes. The POA status vocabulary comprised 6 categories ('POA_Y', 'POA_W_CLIN', 'POA_U_DOC', 'POA_N', 'POA_1_EXEMPT', 'POA_M_MISSING', and 'PAD_POA'). The high fre-

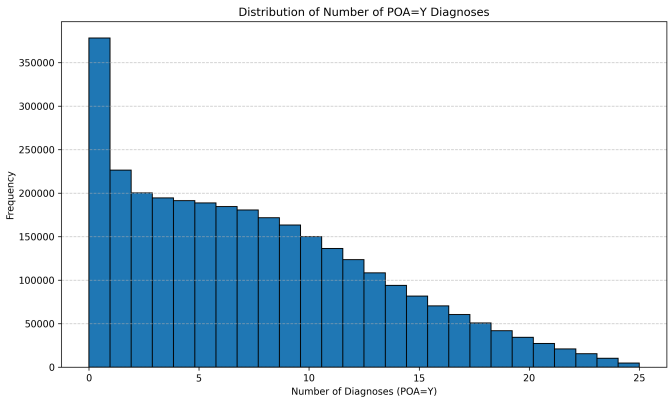

Figure 8. Histogram showing the frequency distribution of the number of diagnoses marked Present On Admission (POA='Y') per patient. This feature reflects the burden of conditions present at admission and was found to be a consistently strong predictor in the models.

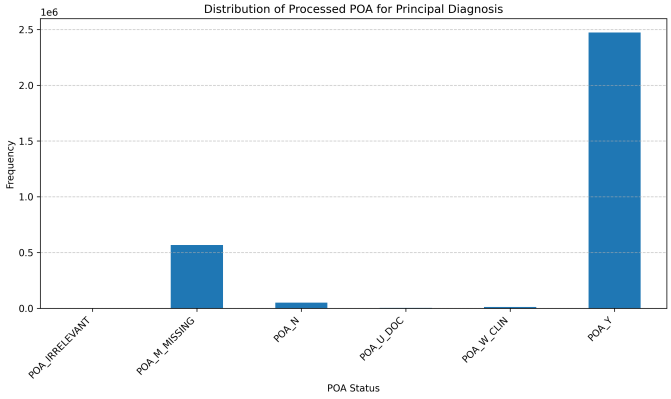

Figure 9. Bar plot showing the frequency distribution of processed Present On Admission (POA) statuses for the principal diagnosis. The distribution highlights the high prevalence of 'Yes' (POA_Y) and 'Missing' (POA_M_MISSING) statuses, indicating a substantial amount of unspecified POA data in the dataset.

quency of 'POA_M_MISSING' underscored the data quality challenge related to POA reporting, as shown in Figure 11. The frequency of the top 30 diagnosis codes in the tokenized sequences is presented in Figure 10.

3.3. *Diagnosis attention encoder training and embedding generation*

A Transformer encoder was trained on the 1% data subsample to learn 160-dimension patient diagnosis representations. The training used in-hospital mortality prediction as a proxy task to guide the learning process, optimizing for validation PR-AUC over 3 epochs. The training and validation loss curves are shown in Figure 12, indicating decreasing loss over training. The model achieved a validation PR-AUC of 0.3728 and

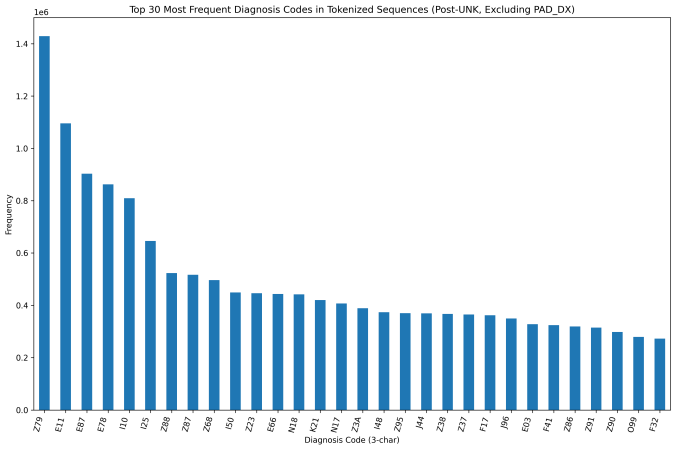

Figure 10. Frequency of the top 30 most common 3-character diagnosis codes in the tokenized sequences used for the diagnosis attention encoder, excluding padding. This distribution reveals the most prevalent conditions present in the dataset.

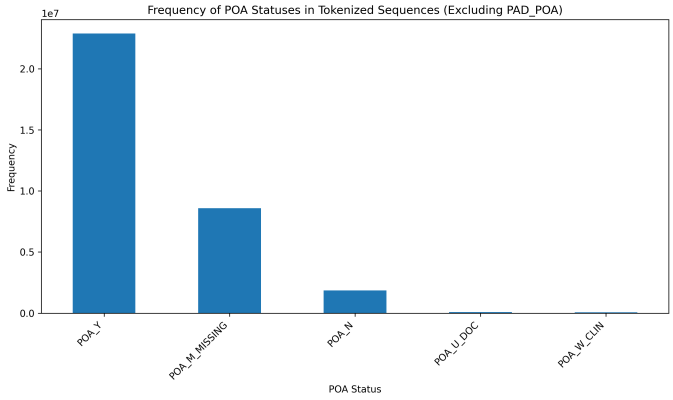

Figure 11. Frequency distribution of Present On Admission (POA) statuses in the tokenized diagnosis sequences (excluding padding). The plot shows that 'POA_Y' (Yes) is the most frequent status, followed by 'POA_M_MISSING', highlighting the significant number of diagnoses with unspecified POA status.

AUC-ROC of 0.9552 on this proxy task, suggesting it learned some predictive signal from the diagnosis-POA sequences. The validation AUC-ROC and PR-AUC over epochs are shown in Figure 13. The trained encoder was then used to generate embeddings for all patients in the subsampled dataset. A t-SNE visualization of test set embeddings showed some degree of clustering by mortality status, further indicating that the embeddings captured relevant information, although separation was not perfect.

3.4. *Predictive model performance*

The learned patient embeddings were concatenated with the processed non-diagnosis features to form the

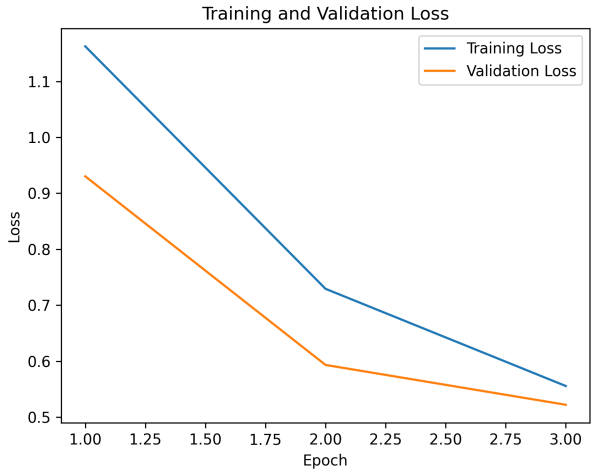

Figure 12. Training and validation loss curves for the diagnosis attention encoder during training for 3 epochs, showing decreasing loss over time.

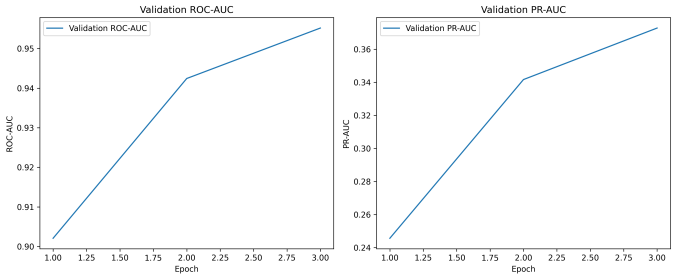

Figure 13. Validation ROC-AUC and PR-AUC for the diagnosis attention encoder trained on a proxy in-hospital mortality prediction task over 3 epochs. Both metrics increased during training, indicating improving performance on the validation set.

"Attention" feature set. This was compared against two baseline feature sets: "Baseline 1 (NonDiag)" using only non-diagnostic features, and "Baseline 2 (SimpleDiag)" combining non-diagnostic features with one-hot encoded top 200 admitting diagnoses and multi-hot encoded top 200 POA='Y' diagnoses. Logistic Regression (LR) and XGBoost (XGB) models were trained and evaluated on the held-out test set for both outcomes.

For In-Hospital Mortality prediction, where the class is highly imbalanced, Area Under the Precision-Recall Curve (AUC-PR) was the primary metric. The XGBoost model trained on the Baseline 2 feature set achieved the highest AUC-PR of 0.266. The Logistic Regression model with Baseline 2 features also performed strongly (AUC-PR 0.228). In stark contrast, the Attention models performed poorly, with XGBoost achieving an AUC-PR of only 0.036 and Logistic Regression 0.056. The Baseline 1 models, using only non-diagnostic features, performed better than the Attention models but worse than Baseline 2, with XGBoost achieving an AUC-PR of 0.072. These results indicate that the simpler, explicit encoding of diagnosis information (particularly POA='Y' diagnoses) provided significantly more predictive power for mortality than the learned attention-based embeddings in this setting.

For Prolonged Length of Stay (PLOS) prediction, the Baseline 2 models again demonstrated superior performance. The XGBoost model with Baseline 2 features achieved the highest AUC-ROC of 0.853 and AUC-PR of 0.631. The Logistic Regression model with Baseline 2 features also performed well (AUC-ROC 0.842, AUC-PR 0.616). The Attention model with XGBoost achieved an AUC-ROC of 0.811 and AUC-PR of 0.567, outperforming the Baseline 1 models (XGBoost AUC-ROC 0.817, AUC-PR 0.577) on AUC-PR but not AUC-ROC. The Logistic Regression Attention model performed less well (AUC-ROC 0.775, AUC-PR 0.502). While the Attention model showed a slight improvement over the non-diagnosis-only baseline for PLOS (in terms of AUC-PR), it still fell short of the performance achieved by the Baseline 2 model using simpler diagnosis features. The ROC and Precision-Recall curves for all PLOS models are shown in Figure 14. A confusion matrix for the best-performing PLOS model (XGBoost with Baseline 2 features) on the test set is provided in Figure 15.

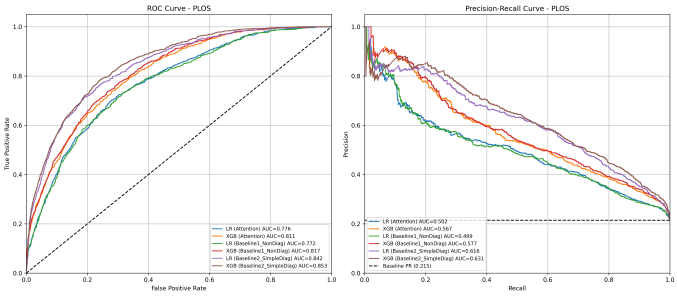

Figure 14. ROC and Precision-Recall curves for Prolonged Length of Stay (PLOS) prediction models on the test set. The plot on the left shows the Receiver Operating Characteristic (ROC) curves, and the plot on the right shows the Precision-Recall (PR) curves. Different lines represent Logistic Regression (LR) and XGBoost (XGB) models using the Attention, Baseline 1 (NonDiag), and Baseline 2 (SimpleDiag) feature sets. Baseline 2 models, incorporating simple diagnosis encoding, achieved the highest AUC-ROC (XGB: 0.853) and AUC-PR (XGB: 0.631), outperforming the Attention models.

Across both outcomes, the models incorporating the simpler, explicit diagnosis features (Baseline 2) consistently outperformed both the models using only non-diagnostic features (Baseline 1) and the models incorpo-

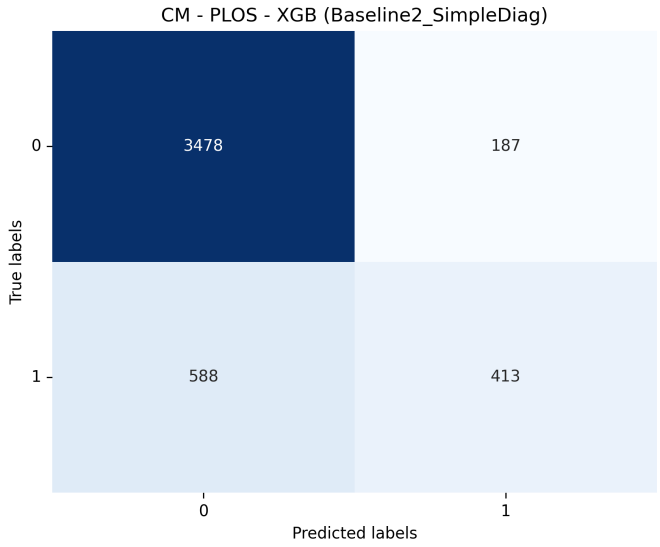

Figure 15. Confusion matrix for the XGBoost model with Baseline 2 (SimpleDiag) features predicting prolonged length of stay (PLOS) on the test set. Values represent counts of true negatives (3478), false positives (187), false negatives (588), and true positives (413).

rating the learned attention-based diagnosis representations.

3.5. *Model interpretation*

SHapley Additive exPlanations (SHAP) analysis was applied to the XGBoost models for both outcomes and feature sets to understand feature importance.

For In-Hospital Mortality prediction, the count of POA='Y' diagnoses (`NUM_POA_Y_DIAGNOSES_scaled`) was a highly influential feature in both the Attention and Baseline 2 models. Geographic factors (`PAT_COUNTY_cleaned`) and patient age were also important across models. Figure 16 shows the mean absolute SHAP values for the Attention model predicting mortality, while Figure 17 provides a beeswarm plot illustrating the direction and magnitude of feature impacts for this model. In the Baseline 2 model, specific diagnosis codes explicitly marked as POA='Y' (e.g., respiratory failure, fluid/electrolyte disorders) were identified as strong predictors, demonstrating the direct value of this specific POA status information. In the Attention model, while the learned embeddings contributed, their overall importance was lower than these prominent non-diagnostic or explicitly encoded diagnosis features.

Similarly, for PLOS prediction, `NUM_POA_Y_DIAGNOSES_scaled` was the most important feature in both the Attention and Baseline 2 models. Type and source of admission, age, and sex were also significant predictors. Figure 18 and Figure 19 show the SHAP results for the Attention model predicting PLOS, while Figure 20 and Figure 21 present the SHAP analysis for the Baseline 2 model. The Baseline 2 model highlighted specific admitting diagnoses (e.g., normal pregnancy supervision, negatively associated with PLOS) and POA='Y' codes (e.g., nicotine dependence, lipid metabolism disorders) as important predictors, reinforcing the value of directly encoding these features. Again, the learned embedding dimensions in the Attention model had lower overall importance compared to these features.

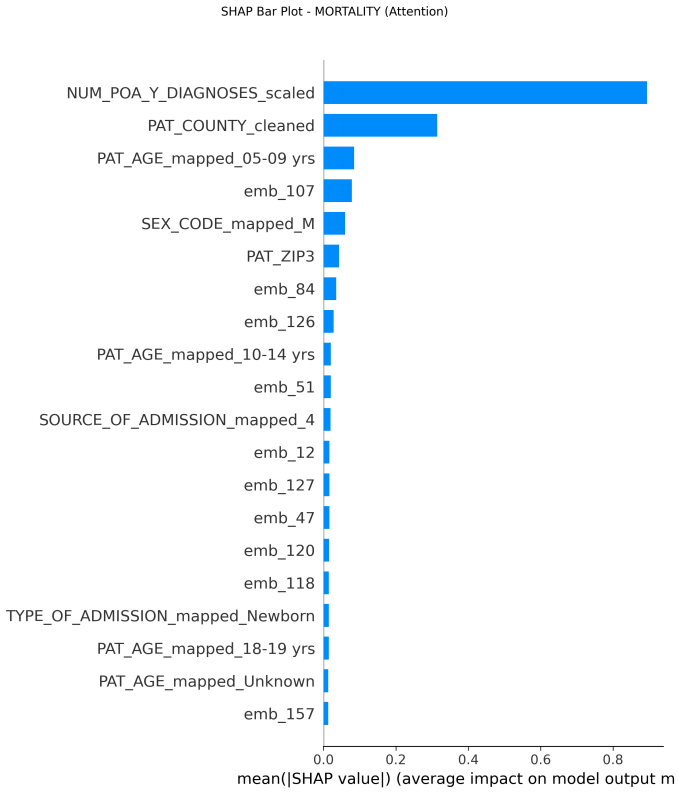

Figure 16. SHAP bar plot showing the mean absolute SHAP values for the XGBoost model predicting in-hospital mortality using the Attention feature set. Features are ranked by their average impact on the model output magnitude. The number of diagnoses present on admission (POA='Y'), patient county, and age categories, along with some learned embedding dimensions, had the largest average impact on mortality predictions.

The consistent high importance of `NUM_POA_Y_DIAGNOSES_scaled` across models and outcomes underscores the clinical significance of the total burden of conditions present at admission as a predictor for both mortality and prolonged stay. The target-encoded geographic features also demonstrated notable predictive power, suggesting spatial variations in outcomes potentially related to patient population characteristics or healthcare access.

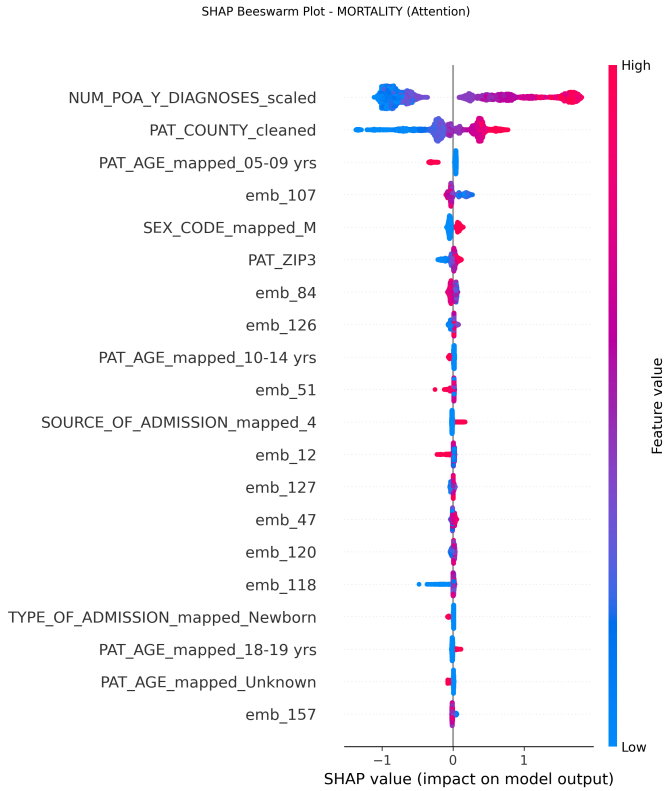

Figure 17. SHAP beeswarm plot showing the impact of features on the XGBoost model output for predicting in-hospital mortality using the Attention feature set. The vertical axis lists features by importance (top to bottom). Each point represents a patient's feature value and its impact (SHAP value) on the model's prediction (horizontal axis). Red indicates high feature values, and blue indicates low feature values. The plot reveals that the scaled number of POA='Y' diagnoses, patient county, specific age groups, and certain dimensions of the learned diagnosis embeddings were the most influential features, with higher values of the number of POA='Y' diagnoses generally increasing predicted mortality risk.

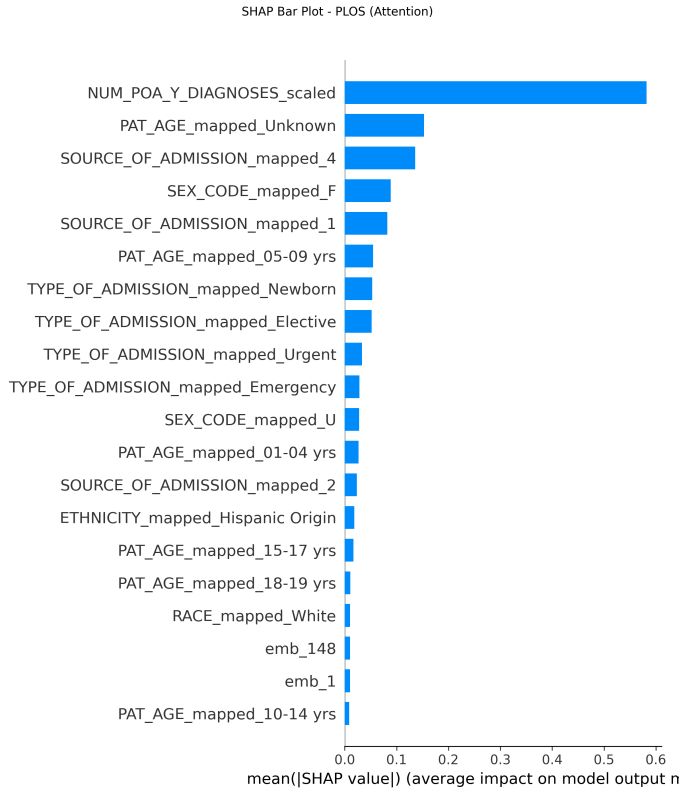

Figure 18. SHAP bar plot showing the average impact of features on the XGBoost model predicting Prolonged Length of Stay using the Attention feature set. Features are ranked by their mean absolute SHAP value. The number of POA='Y' diagnoses, patient age, and admission source/type were the most important features for this model.

An attempt was made to visualize attention weights from the trained Transformer encoder to gain insight into which diagnosis_POA pairs were most attended to, but this was unsuccessful, limiting the ability to interpret the internal workings of the attention mechanism directly.

In summary, the evaluation revealed that while the attention-based encoder learned representations with some predictive signal in a proxy task, models using these embeddings did not surpass the performance of models leveraging simpler, explicit diagnosis feature engineering and non-diagnostic features for the target outcomes. The number of diagnoses present on admission emerged as a consistently strong predictor. The results suggest that for this dataset and task, given the constraints (particularly data subsampling for the Transformer), simpler feature engineering approaches provided more effective inputs for predictive modeling.

4. CONCLUSIONS

In this study, we addressed the critical challenge of predicting in-hospital mortality and prolonged length of stay using administrative discharge data, focusing on leveraging complex diagnosis codes and their Present On Admission (POA) status. We proposed a novel deep learning approach utilizing a Transformer encoder to learn contextualized patient representations from diagnosis-POA sequences, explicitly modeling missing POA information. This learned embedding was then combined with other admission-time features, including demographics, admission type, and an engineered count of diagnoses present on admission, to train predictive models.

Using data from the 2018 Texas Hospital Inpatient Discharge Public Use Data File, we trained Logistic Regression and Gradient Boosting models on the combined

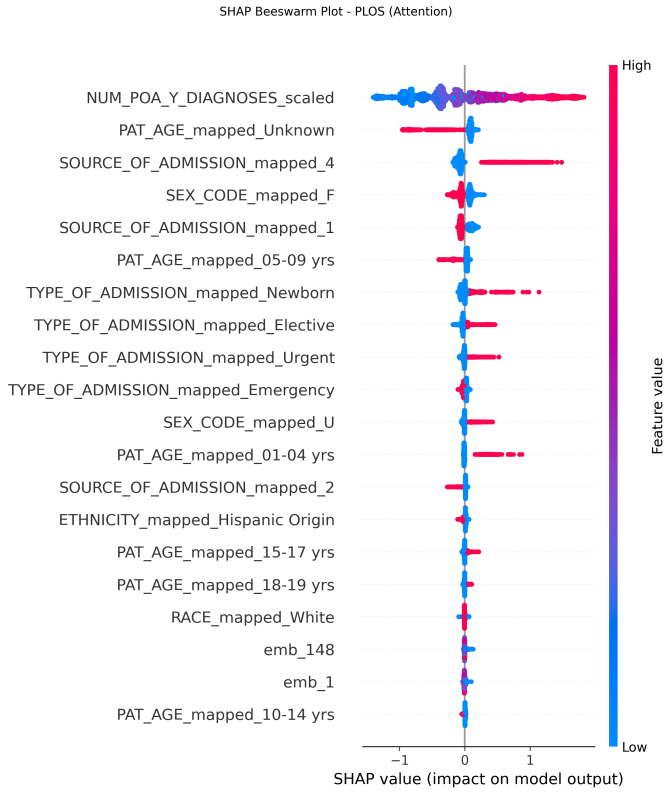

Figure 19. SHAP beeswarm plot showing feature importance for the XGBoost model predicting Prolonged Length of Stay (PLOS) using the Attention feature set. Each point represents a patient, positioned horizontally by the SHAP value for that feature and colored by the feature value (red high, blue low). Features are ordered by overall importance. The plot illustrates that the number of diagnoses present on admission (`NUM_POA_Y_DIAGNOSES_scaled`), patient age (especially unknown age), admission source and type, and sex were the most influential features for this model, while learned embedding dimensions (e.g., `emb_148`, `emb_1`) had lower impact.

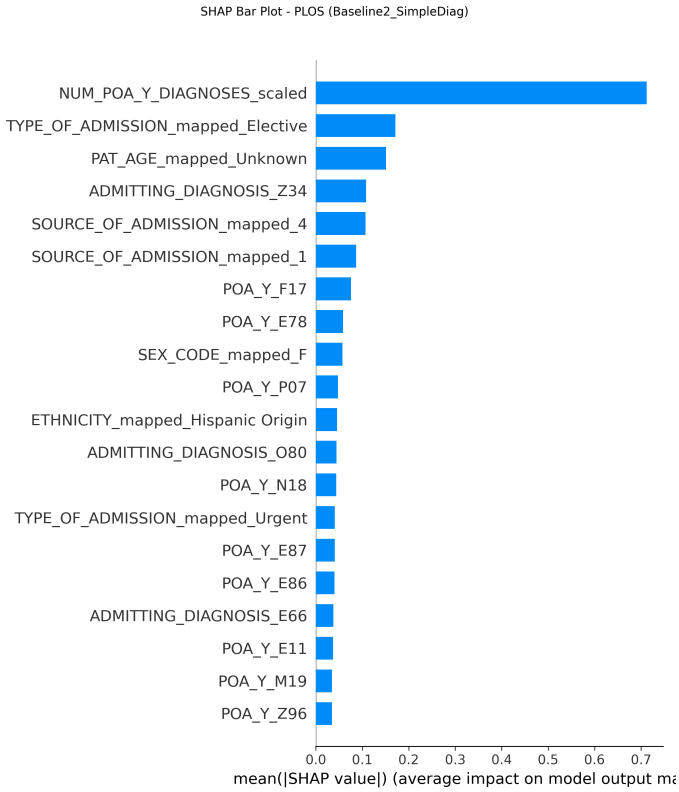

Figure 20. SHAP bar plot showing the mean absolute SHAP value for the top features in the XGBoost model predicting prolonged length of stay using the Baseline 2 (SimpleDiag) feature set. This plot indicates the average impact of each feature on the magnitude of the model's output, highlighting the most influential predictors such as the scaled number of diagnoses present on admission.

feature set and compared their performance against baseline models using only non-diagnostic features or simpler, explicit diagnosis encodings (top N principal and POA='Y' diagnoses). The Transformer encoder was trained on a 1% subsample of the data, showing some ability to learn predictive signals in a proxy mortality prediction task.

However, the evaluation results demonstrated that the final predictive models incorporating the learned attention-based diagnosis representations did not outperform the baseline models, particularly those utilizing a simpler, explicit encoding of diagnosis information alongside non-diagnostic features. For both in-hospital mortality (evaluated by AUC-PR) and prolonged length of stay (evaluated by AUC-ROC and AUC-PR), the models trained on the simpler diagnosis features (Baseline 2) achieved superior performance. The models using only non-diagnostic features (Baseline 1) generally performed better than the attention-based models for mortality, and comparably or slightly worse for PLOS, depending on the metric.

A key finding from the model interpretation using SHAP was the consistent high importance of the engineered feature representing the count of diagnoses explicitly marked as Present On Admission (`NUM_POA_Y_DIAGNOSES`). This underscores the significant predictive power of the patient's condition burden at the time of hospital admission. Furthermore, the baseline models that explicitly included the identities of frequently occurring POA='Y' diagnoses also highlighted their strong predictive value. Other important features included patient age, admission type, and geographic factors.

From these results, we learned that while complex deep learning architectures like the Transformer can theoretically capture intricate patterns in sequential or set-

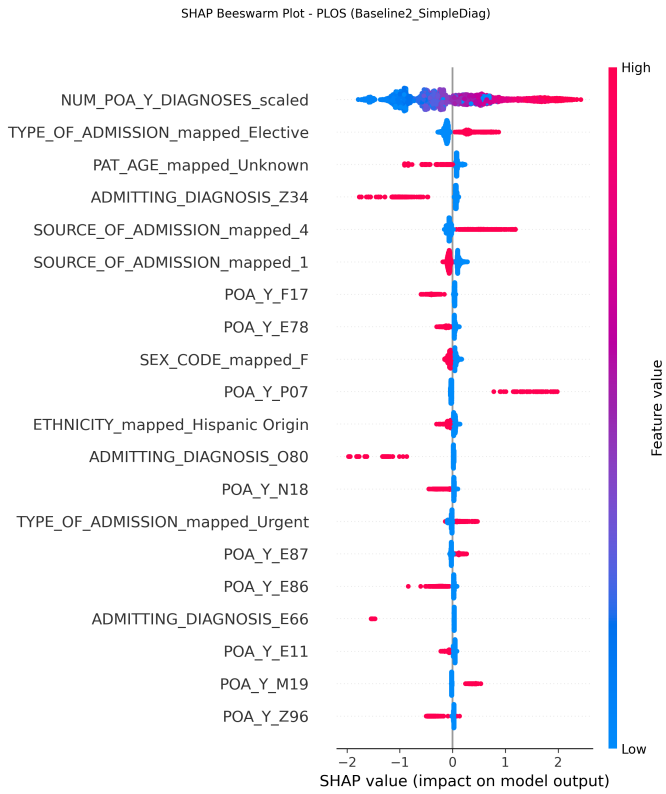

Figure 21. SHAP beeswarm plot showing feature importance for the XGBoost model predicting prolonged length of stay (PLOS) using the Baseline 2 feature set. The plot illustrates how different feature values (color) contribute to the model output (horizontal position), identifying the most influential factors for PLOS prediction.

based data like diagnosis-POA lists, their effectiveness in practice is highly dependent on factors such as the volume of training data available for the representation learning component (limited here by subsampling) and how effectively the learned representations are integrated with other crucial clinical and administrative features. In this specific context and dataset, simpler, more direct feature engineering approaches, particularly those that explicitly leverage the Present On Admission status for specific diagnoses, provided more robust and predictive signals for the target outcomes. The value of the count of diagnoses present on admission as a readily available and powerful predictor is also confirmed. These findings suggest that while exploring advanced representation learning is valuable, well-designed feature engineering remains a strong baseline and can sometimes outperform complex end-to-end learning systems, especially when computational or data volume constraints are present or when the most predictive signals are relatively straightforward to extract explicitly.