Humana-Mays Healthcare Analytics

2023 Case Competition

Osimertinib Therapy Discontinuance: Prediction and Segmentation Analysis

Table of Contents

1. EXECUTIVE SUMMARY

1.1 Study Proposal

Lung Cancer is one of the leading causes of death, and new treatments are coming to the market all the time. However, many of these treatments are associated with potentially significant side effects, which can make it difficult for patients to stay adherent to their life-saving medications. One of these medications is Osimertinib, known to be a largely effective medication, but can cause side effects such as nausea, fatigue, pain, high blood glucose, and constipation. Many of these side effects are manageable with proper counseling and avoidance techniques, but many patients may opt to discontinue their treatment rather than seek guidance on managing them. Approximately one-quarter of Humana members taking Osimertinib have side effects and discontinue their Osimertinib therapy within the first 6 months.

New hope for Lung cancer patients: By identifying patients who are at risk of discontinuing their treatment, Humana can provide them with the support and resources they need to stay adherent to their medication and improve their chances of survival. This study has the potential to make a significant impact on the lives of small-cell lung cancer patients and their families.

Our objective: Our objective is to assist Humana in identifying members who are most likely to discontinue therapy that ends prematurely (before 180 days) and have an ADE reported at some time during the therapy by applying big data analysis and machine learning methods.

1.2 Modeling

To achieve the best performance of modeling, we carried out comprehensive studies to understand the dataset features and the business issue. First, we built a predictive model to identify members who discontinued the therapy before 180 days and had reported ADE during the therapy. We performed exploratory data

analysis to understand the nuances of the data, which allowed us to do feature engineering and data preparation.

Post that, we employed the Gini Index, XGBoost, and LightGbm to perform feature importance. Based on the three models' intersections, we developed a better understanding of the most important features in our model. Then we applied LightGBM along with hyperparameter tuning to do preliminary prediction and compared their performances and corresponding AUC. Finally, we got the best performance with an AUC of 0.958 with LightGBM for the training dataset. Further analysis and recommendations on predicting patients who discontinue therapy are based on the features and insights we derived.

1.3 Recommendation

Based on our insights, we propose Humana adopts a customized strategy to enhance Osimertinib adherence among its members, delineated into three segments predicated on economic conditions, patients' health status, and adherence metrics:

Segment 1: Provision of Financial Support and Facilitated Home Care

Segment 2: Medication Reminder and Monitoring Services

Segment 3: Tailored Medication Management and Monitoring

Segment 4 (All Members): Subscription-based model for medication access ensuring continuous medication intake, along with Medication Administration Reminder and Monitoring features.

Through the deployment of this nuanced approach, Humana can address the diverse challenges patients encounter in maintaining adherence.

1.4 Cost Analysis:

Post segmenting our patient characteritics, we evaluated the financial implications of our recommendations and provided insights on whether the overall cost of implementing these data-driven and actionable suggestions could be offset or remain lower than Humana's current expenses.

2. CASE BACKGROUND

2.1 Context

In recent years, there have been significant advancements in oncology research and the development of new therapies, yet lung cancer remains a leading cause of death worldwide. According to the American Cancer Society, Lung cancer is by far the leading cause of cancer death in the US, accounting for about 1 in 5 of all cancer deaths. Each year, more people die of lung cancer than of colon, breast, and prostate cancers combined. It is the second most common cancer in the United States, after skin cancer.

Non-small cell lung cancer (NSCLC), the most common type of lung cancer, tends to recur when diagnosed at advanced stages, which makes treatment challenging. Osimertinib is very effective in treating this type of lung cancer, with a response rate of over 70%. As per Yale School of Medicine, of the 682 patients with stage IB-IIIA NSCLC enrolled in the trial, 88% of patients treated with osimertinib following surgery were still alive five years later. It is a third-generation epidermal growth factor receptor tyrosine kinase inhibitor.

Despite the *effectiveness of Osimertinib*, in treating specific types of lung cancer, patients often face challenges related to treatment tolerability and side effects. The most common side effects include diarrhea, rash, musculoskeletal pain, dry skin, skin inflammation around nails, sore mouth, fatigue, and cough.

2.2 Problem Statement

The side effects of Osimertinib can be severe in some cases, and they can lead to patients discontinuing therapy prematurely, impacting patient outcomes. We are committed to addressing these healthcare challenges by leveraging data analysis and machine learning techniques. The purpose of this analysis is to help Humana to identify its members who are at risk of discontinuing therapy prematurely and experiencing adverse drug events during their treatment course, with the ultimate goal of improving patient care and outcomes through data-driven recommendations.

To achieve this goal, we first applied a classification model to predict which members are most likely to experience osimertinib discontinuation and reported ADE based on the provided data. Then we identified the most important features affecting discontinuation, categorized the members into three major groups, and proposed potential solutions to address their non-adherence problems.

3. DATA ANALYSIS

3.1 Dataset Description

Humana supplied a total of six datasets comprising member healthcare information. These datasets are divided into two categories: training datasets, which were employed for model training, and holdout datasets, which are reserved for predictions. Within each category, there are three distinct datasets, including medical claims, pharmacy claims, and a target dataset.

- **Medical claims**: medclms_train"(100159 records) and "medclms_holdout" (23232 records) datasets contain all medical claims for an individual within the 90 days preceding their Osimertinib therapy and extending through the therapy period. There are 536 unique patients in the medical dataset. These datasets encompass visit and process dates, diagnosis codes, and indicators for diagnosis codes of interest.

- **Pharmacy claims**: "rxclms_train" (32133 records) and "rxclms_holdout"(6670 records) are two datasets that contain simplified information about all pharmacy claims for an individual during the 90 days before their Osimertinib therapy and throughout the therapy duration. There are 1160 unique patients in the pharmacy dataset. This data comprises service dates, process dates, drug identifier codes (NDC), and indicators for drug codes of interest.

- **Target:** "target_train" (1232 records) and "target_holdout" (420 records) datasets contain information about the therapy start and end dates, the target identifier, and protected attributes for each individual. The target variable, 'tgt_ade_dc_ind,' ==1 when the patient concludes the therapy prematurely (before 180 days) and has an Adverse Drug Event reported during the therapy period. There are 1232 unique patients in the target train dataset, and the targeted group has 117 members, which accounts for about 9.5% of all the patients.

The following sections provide a brief overview of the available data. Detailed descriptions of all fields are available in data_dictionary.csv.

Group (# of variables)	Prefix/Name	Feature Description	Data type
Medical Claims	'therapy_id', 'medclm_key', 'clm_unique_key '	Unique identifiers for patient and medical claims	Categorical
	' primary_diag_cd', 'diag_cd#',	Diagnosis codes for a medical claim	Categorical
	'visit_date','process_date'	Date that this claim was visited and processed	Datetime
	'pot', 'util_cat', 'hedis_pot'	Place of treatment for this claim with different granularity.	Categorical
	'ade_diagnosis', 'seizure_diagnosis', 'pain_diagnosis', 'fatigue_diagnosis', 'nausea_diagnosis', 'hyperglycemia_diagnosis', 'constipation_diagnosis', 'diarrhea_diagnosis'	Indicates if the diagnosis codes in this claim report certain side effect	Categorical (binary integer)
Pharmacy Claim	'therapy_id', 'document_key'	Unique identifiers for patient and pharmacy claims	Categorical
	'service_date', 'process_date'	Date when the claim was serviced and processed	Datetime
	'rx_cost', 'tot_drug_cost_accum_amt'	Singal cost and acculated cost of the claims	Numeric (float)
	'Ndc_id',	Provides detailed information about the medication, including its ndc id, class and quantity served.	Categorical
	'gpi_drug_group_desc', 'gpi_drug_class_desc', 'hum_drug_class_desc', 'strength_meas', 'metric_strength'		
	'specialty_ind', 'clm_type', 'ddi_ind', 'anticoag_ind', 'diarrhea_treat_ind', 'nausea_treat_ind', 'seizure_treat_ind'	Indicates if the claim is for a drug used to treat certain side effect	Categorical (binary integer)
	'pay_day_supply_cnt	The number of days supply of a drug	Numeric (integer)
Target	'id', 'therapy_id'	Unique identifiers for patient	Categorical
	'therapy_start_date', 'therapy_end_date'	Date when the patient started and end therapy	Datetime
	'tgt_ade_dc_ind'	Indicate if the patients end therapy before 180 days and reported ADE	Categorical (binary integer)
	"sex_cd', 'cms_disabled_ind', 'cms_low_income_ind'	Personal attributes include gender, physical and financial status of patients.	Categorical
	'est_age'	Indicate the ages of patients.	Numeric (integer)
	'race_cd'	Indicate the race of patients (7 races)	Categorical (muti-class integer)

3.2 Descriptive Statistics

Humana's dataset provides a wealth of membership information, and before the data cleaning, we obtained a brief understanding of the nature of members in the Humana database.

- **Age Distribution:** Humana's database contains members' age information, in which the average age of members is 73, and the 25% and 75% quantiles are 68 and 80 years old, respectively. We can also observe that most of Humana's members are 70-79 years old.

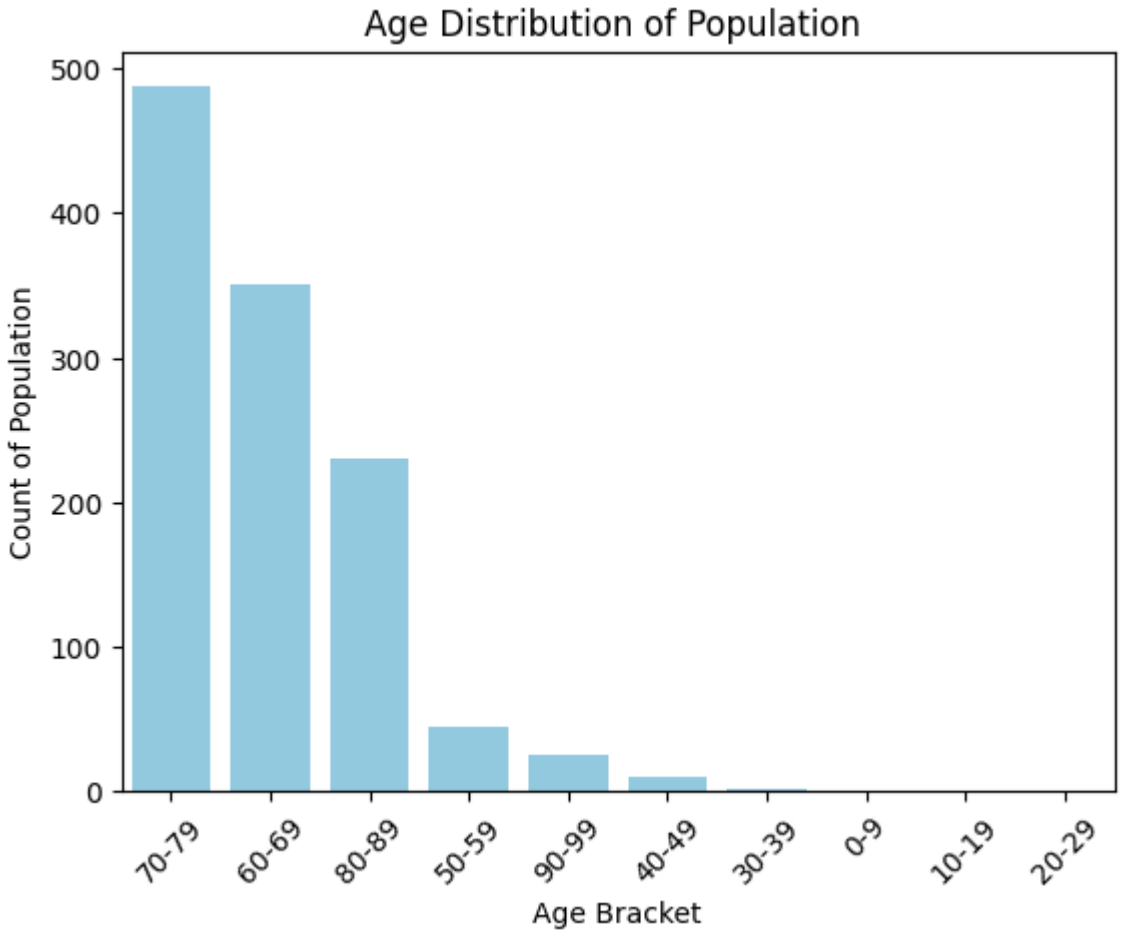


Figure 1: Age Distribution of Patients

- **Gender:** Most of the records consisted of women-identifying members with 815 counts and 334 male-identifying population.

- **Race:** In the ‘target_train’database, feature ‘race_cd’ records the racial information of members, including Whites, Blacks, Asians, Hispanics, Native Americans, Others, and Unknown. Upon analysis, the largest proportion is the white group, accounting for 56% of the population, and other/Native American categories account for the least at 3% of the population.

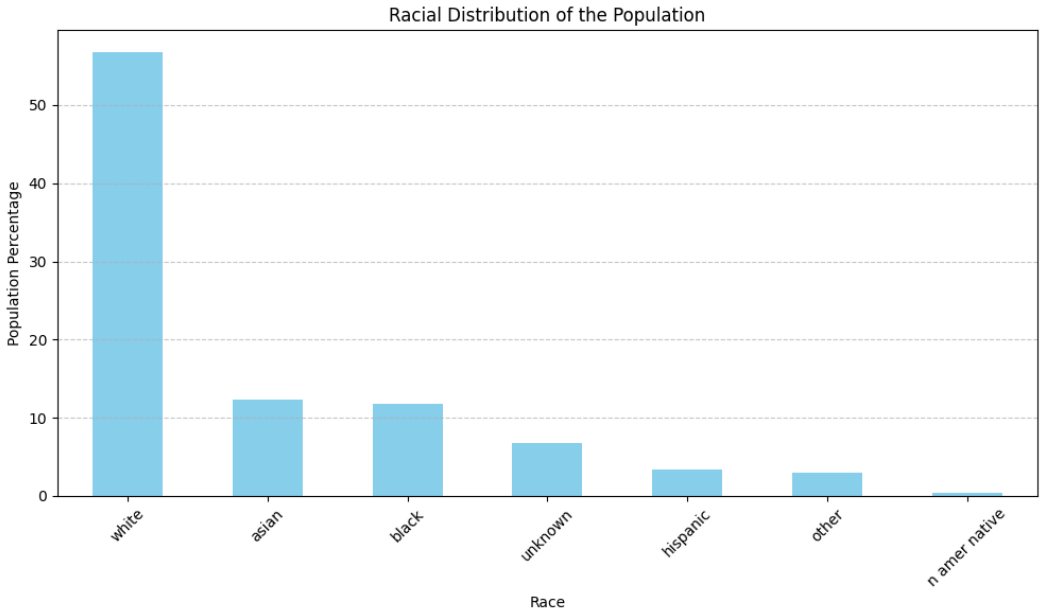


Figure 2: Race Segregation of Patients

3.2.1 Exploratory Data Analysis

- **Most common side effects amongst those who discontinued therapy:** In the medical_train dataset, among the adverse events (ADEs) that patients have claimed, fatigue is the most frequently claimed disease, followed by constipation and nausea

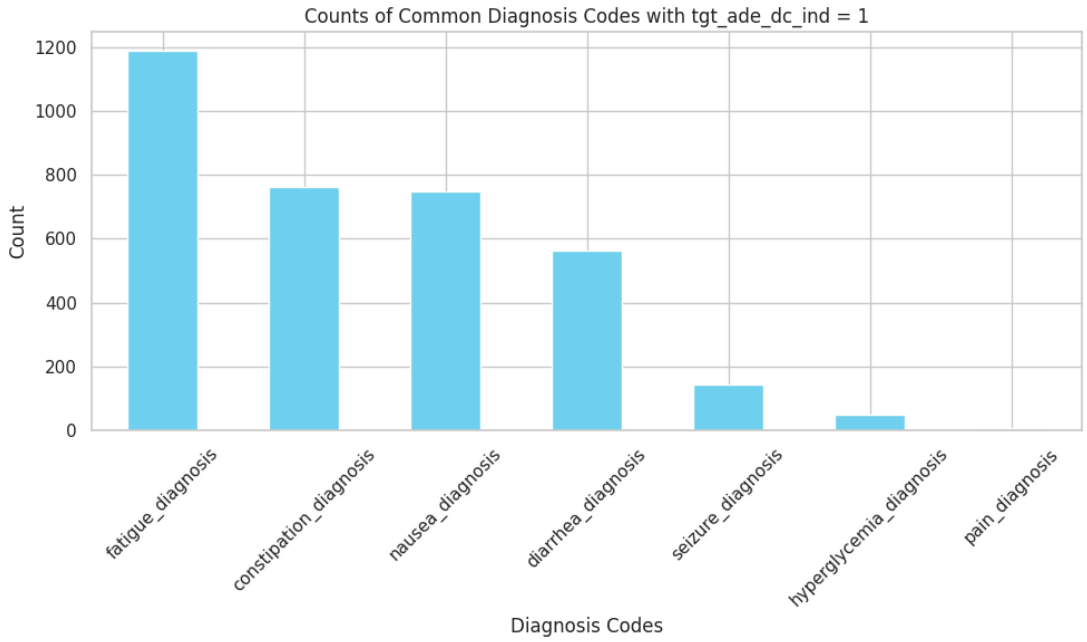


Figure 3: Count of Diagnosis for Patients whose Treatment was Unsuccessful

- **Place Of Treatment (POT):** The distribution of places of treatment shows that 43% of patients get diagnosed in an “unknown” location, followed by “Outpatient” at 38%. These two POTs account for the most common place of treatments amongst the patients.

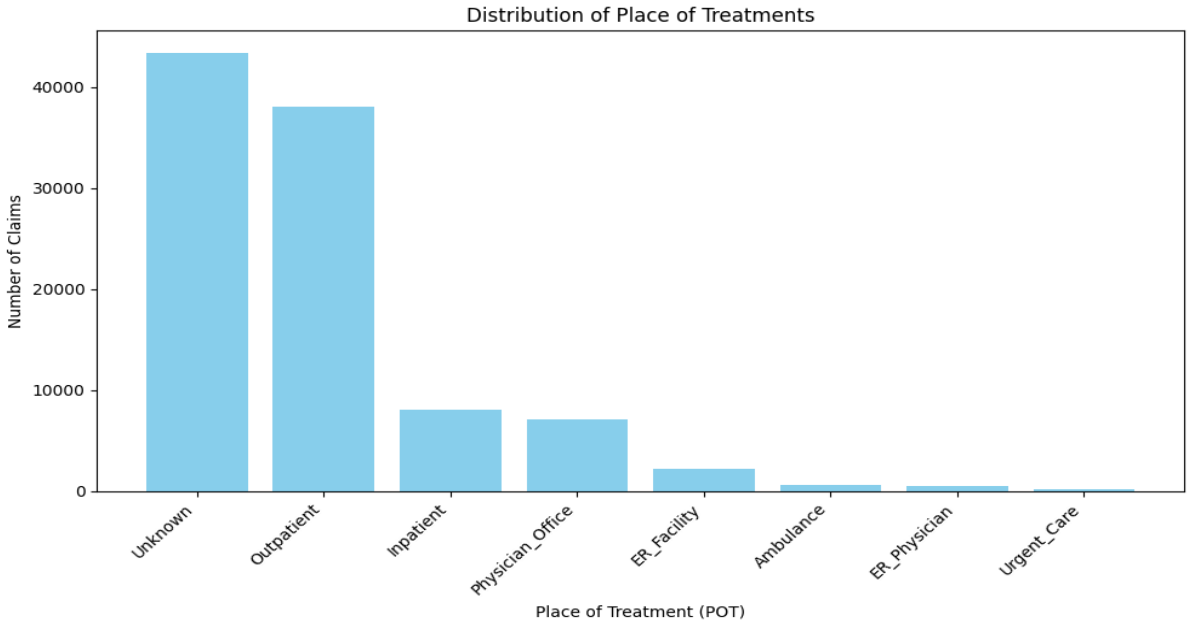


Figure 4: Distribution of Place of Treatment Among Patients

- **Therapy Duration & Pay Day Supply:** We calculated therapy duration in the training dataset by subtracting therapy start dates from end dates. An interesting observation emerged when we analyzed patients who discontinued treatment before 180 days and reported adverse drug events.

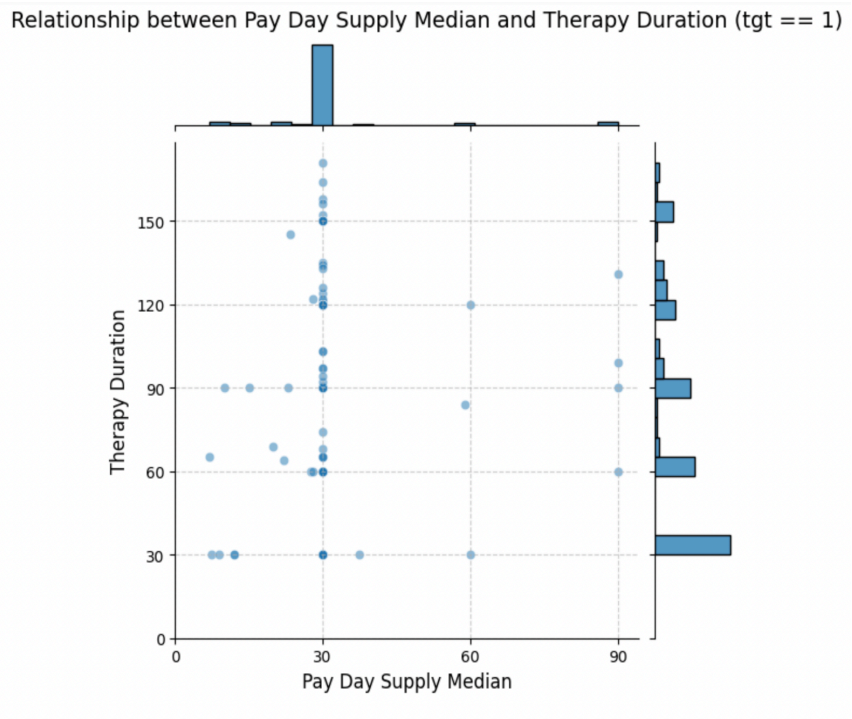


Figure 5: Relationship Between Pay Day Supply of the Medicine and Therapy Duration Among Patients with Unsuccessful Treatment

The histogram on the right side shows therapy duration, revealing that a significant number of patients stopped treatment precisely at the end of the first 30 days. Distinct spikes appeared at 60 days, 90 days, and so on, with a few patients discontinuing in between. Most patients were prescribed a 30-day supply, matching an Osimertinib package containing 30 tablets, where one tablet is taken daily. This suggests that patients often complete an Osimertinib package before discontinuing treatment. This insight underscores the importance of monitoring patients' medication administration history, with further recommendations provided in the following sections.

3.3 Data Cleaning and Imputation

3.3.1 Drop Duplicate Records and Irrelevant Columns

Upon inspecting the dataset, we uncovered duplicate records in both the medical claim and pharmacy claim datasets. Notably, individual patients had submitted multiple medical claims, each of which had a unique 'medclm_key,' though some of the claims contained identical claim information.

To enhance our analysis, we chose to eliminate these duplicates. Consequently, we removed 66,051 records from the medical training dataset and 15,008 records from the medical holdout dataset. This reduced the medical training dataset from 100,159 to 34,108 records and the medical holdout dataset from 23,232 to 8,224 records. In the pharmacy dataset, we identified only one duplicate record, resulting in a reduction from 32,133 to 32,132 records. We conducted data exploration and identified trends in the dataset after the duplicate removal process.

We dropped columns such as ‘medclm_key’, ‘clm_unique_key’, ’ process_date’, ‘reversal_ind’, ‘clm_type’, ‘hedis_pot’ etc as these features were irrelevant in performing the analysis directed towards the prediction problem.

3.3.2 Data Types Transformation

The first step of our data cleaning is to regulate the data types of all features. We found that some categorical features, such as ‘cms_race_cd’, have inconsistent data types, that is, some features are of numerical type, and some are of categorical type. So we first performed type-consistent conversion on all categorical variables. Next, we converted all missing values to NaN to avoid numerical data being recognized as categorical data in the following data processing process and also prepared the data for imputation of missing values.

3.3.3 Missing Value Imputation

Before we formally address missing values, it is essential to gain an overall understanding of the situation regarding missing values in the dataset. Since not all of the 1232 patients in the training dataset have both medical and pharmacy claim data, there will be missing values when we merge the three datasets. More than 15 columns have over 50% missing values, while 9 columns have less than 5% missing values. Depending on the feature type, we applied relevant missing data imputation methods.

For categorical columns, most of them are transformed into numerical variables during our feature engineering process to implement machine learning models (mentioned in detail below). Each category under a categorical variable is transformed into a separate column, and the values represent the count of that

specific category appearing in the dataset. After this transformation, we will replace the numeric missing values with the median.

3.4 Feature Engineering

One significant challenge is dealing with multiple claims made by the same patient. Without proper feature engineering to address this issue, the dataset may contain duplicate information, which could adversely impact our model's performance.Therefore, we decided to employ feature engineering techniques to consolidate each patient's multiple medical and pharmacy claim records. After this process, each patient will have only a single record that provides a comprehensive overview of their information from their multiple claims.

3.4.1 Transforming Categorical Variables into Numeric Variables

1. Calculate Proportions for Binary Variables

First, we converted the categorical variables into a numerical format. For patients with multiple claims of the same category, we calculated the proportions of this category among all their claims. For example, the 'Y_mail_order_proportion' quantifies the proportion of a patient's mail-order claims.

If a patient has a total of 10 pharmacy claims, and 5 of them are for mail order ('Y' value), the proportion for that patient is 0.50. We transformed 12 binary variables into a numeric format using this method. These variables include 'ade_diagnosis,' 'seizure_diagnosis,' 'pain_diagnosis,' 'fatigue_diagnosis,' 'nausea_diagnosis,' 'hyperglycemia_diagnosis,' 'constipation_diagnosis,' 'diarrhea_diagnosis,' 'mail_order_ind,' 'generic_ind,' 'maint_ind,' and 'specialty_ind.' This transformation facilitated their integration into our analysis.

2. Count Multiple-class Categorical Variables

- *Count of the 'Diagnosis Letter' Feature*

In our medical dataset, we have more than 1000 unique diagnosis codes. While we value this information, we are cautious not to introduce an excessive number of

variables through one-hot encoding, as it can potentially slow down or impair our model's performance. To address this, we conducted research and discovered that the first character of a diagnosis code, an uppercase letter, represents a broad category or chapter of diseases or conditions. We then extracted the first-letter information from all the diagnosis code columns and aggregated it on an individual level. As a result, each unique therapy_id now has its respective count of 24 different diagnosis letters.

- *Count of POT (Place of Treatment)*

The dataset included three columns containing information about the location of treatment received by patients. To aggregate them on a personal level, we checked the count of visits made by each unique ‘therapy_id’ in each place of treatment. For example, we calculated how many times a patient visited a Physician_Office or received 'Outpatient' treatment, which is stored in the corresponding column.

3.4.2 Aggregating Claim History

1. Count of Medical and Pharmacy Claims

To gain insights into patients' claim history, we introduced two key variables: 'number_of_med_claims' and 'number_of_rx_claims.' These variables provide a straightforward numerical representation of a patient's total medical and pharmacy claims.

2. Obtain the Final Spending for Each Patient

Furthermore, we addressed the variable 'tot_drug_cost_accum_amt,' which records incremental values for patients across new records. To calculate the patients' total spending throughout their therapy duration, we created 'tot_drug_cost_at_end' by aggregating these values.

3. Averaging for Multiple Claims

We also calculated the gap days between the date the patient received a service and the date the claim was processed for each claim. Subsequently, we computed the average gap day for each patient's claims processed by Humana and created variables 'average_service_process_gap' in the medical dataset and 'average_process_gap' in the pharmacy dataset. Finally, for each patient's claims with different supply days for their medication, we created the variable 'average_pay_day_supply' to calculate the average number of days patients typically claim for their supply.

By applying these feature engineering techniques, we have effectively prepared our dataset for in-depth analysis and insights.

3.5 Feature Selection

After aggregating and performing feature engineering on both the medical and pharmacy datasets, we obtained a combination of original columns and new columns created to consolidate the information. In our feature selection process, we initially dropped some of the original columns that had already been aggregated. For example, we removed 'ade_diagnosis' since it had been converted into 'ade_diagnosis_proportion.'

Next, we utilized feature importance analysis to identify the top features contributing to the prediction of the target variable. Ultimately, after this feature selection process, we included 41 columns from the medical dataset and 13 columns from the pharmacy dataset.

3.6 Merged Dataset

After conducting feature engineering and feature selection on both the medical and pharmacy datasets, we proceeded to merge them with the target dataset for subsequent modeling. We used the therapy ID as the key identifier for merging. Before merging the datasets, we randomly selected a single unique record for each patient, identified by their unique therapy ID. Since we had already aggregated information for each patient, these records now included comprehensive details from multiple records associated with each patient.

Following this selection process, our medical dataset contained 536 records, and the pharmacy dataset contained 1,160 records. Upon merging these two datasets with the target dataset, our final merged dataset comprised a total of 1,232 records and a total of 62 features for our modeling.

4. MODELING

4.1 Model Selection

After thorough data preparation and effective dataset aggregation, the key determinant of our prediction accuracy was the model selection. Our primary objective is to predict which members were most likely to discontinue therapy and experience adverse effects. We initially framed this as a classification prediction task. Upon aggregation, we were left with a dataset containing a single record for each patient, resulting in 1232 rows and 62 features. We conducted experiments with various models, including logistic classifier, XGBoost, and LightGBM, to perform preliminary predictions and subsequently compared their performance.

In the end, after extensive testing, LightGBM emerged as the top-performing model with an impressive AUC score of 0.958 for the training dataset. Logistic regression exhibited an AUC of 0.755, while XGBoost delivered a respectable score of 0.896.

4.2 Final Model Construction

Based on the AUC metric shown in the figure below, the LightGBM Classifier achieved an AUC score of 0.958, surpassing the other models. Consequently, we opted to utilize LightGBM for predictions on our holdout dataset. In addition to its exceptional predictive performance and rapid processing speed, LightGBM excels in addressing the imbalanced dataset issue we encountered. Out of 1232 observations, only 9.5% of members (117) had unsuccessful outcomes, while the majority achieved success with no reported adverse effects, transfers, or mortality during therapy.

To conduct a more in-depth analysis of our model's performance, we also computed the confusion matrix. When we set the threshold to 0.5, the true positive rate of predictions reached 91.3%, while the false positive rate was a relatively low 10.71%. This demonstrates the model's excellent performance.

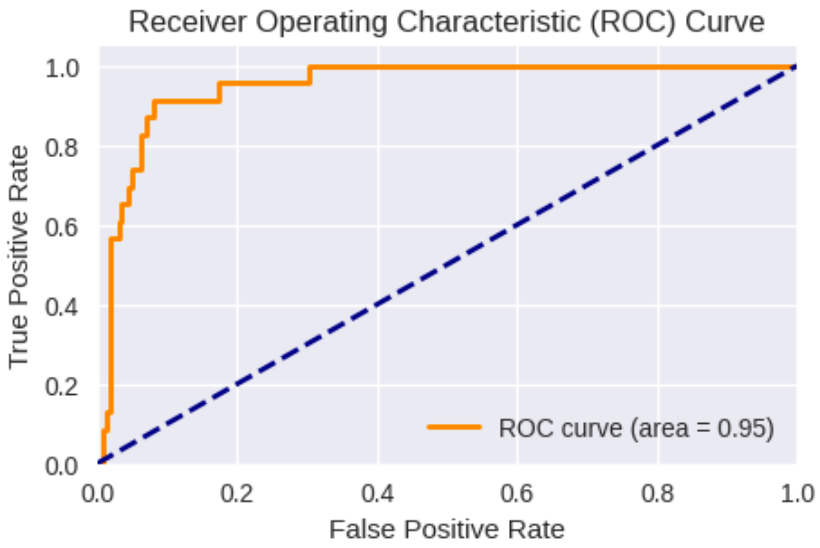


Figure 6: ROC Curve

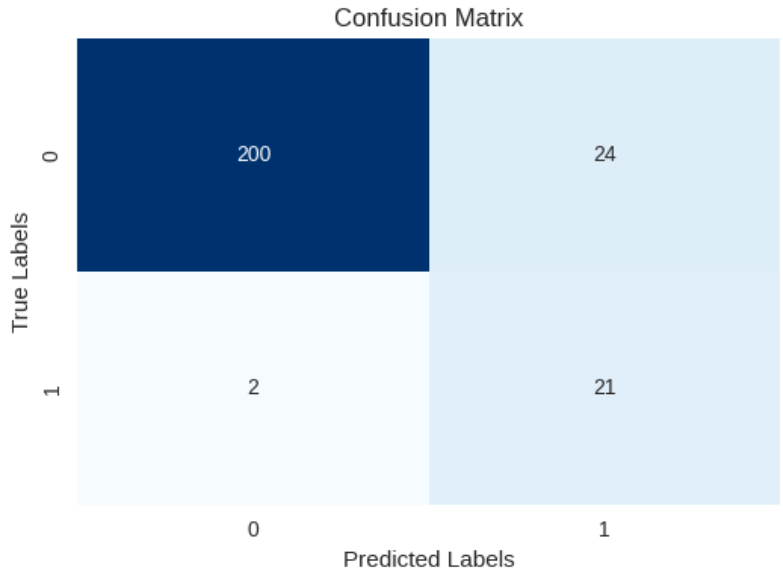


Figure 7: Confusion Matrix

After tuning the hyperparameters we use in our model, they are decided as the following:

Parameter	Value	Description
learning_rate	0.0612	The rate at which the model weights are updated after working through each batch of training examples. A lower learning rate is more capable of finding a better optimum.
max_depth	36	The maximum depth for a tree model.
		Overfitting speed increases as the depth increases. Therefore it should be set by taking into consideration the size of the dataset.
min_child_samples	45	The minimum number of data samples required in a leaf node to control tree complexity and overfitting.
n_estimators	100	This is the number of predictors (trees to build) that we want the model to build.
		Increasing the number of trees can improve model performance while resulting a longer training time and a higher chance of overfitting
num_leaves	20	The maximum number of leaves in one tree
reg_alpha	0.347	L1 regularization
reg_lambda	0.449	L2 regularization
scale_pos_weight	10	Balance the class distribution by assigning higher weight to the positive class
early_stopping_round	20	The training will stop if one metric of one validation data point does not improve in the last early_stopping_rounds round.
		We avoid setting it too large to decrease the chance of overfitting.
feature_fraction	0.9755	The percentage of features for our model to randomly select at the beginning of constructing each tree. It is used to reduce the total number of splits that have to be evaluated to add each tree node.

5. KEY PERFORMANCE INDICATOR ANALYSIS

5.1 Feature Importance

To better understand the model and important features, and drive insights from the model. We looked at the top 20 important features in Lightgbm gain importance and the top 20 important SHAP values.

● *Gain Importance*

We used the built-in Lightgbm feature importance function to get the most important features after tuning and training the model. The calculated numerical value of “gain” to take each feature’s contribution to each tree in the model is the most common method to evaluate the importance of the features in the model. The top 20 important features are shown in the following figure.

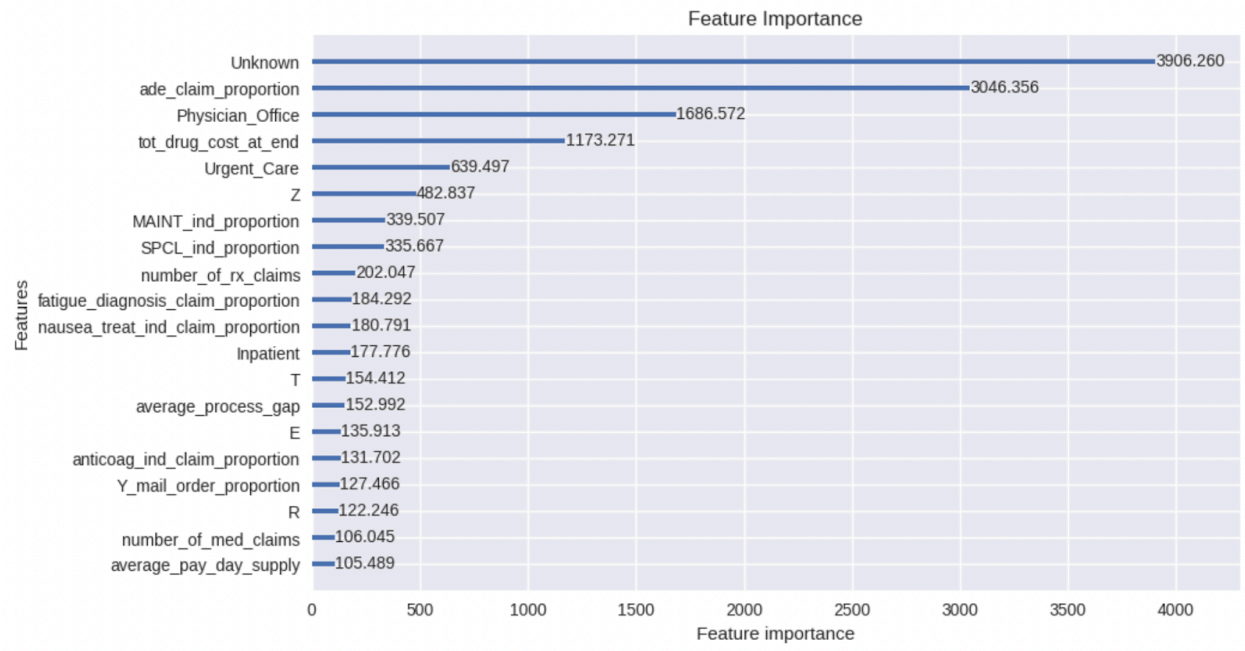


Figure 8: Feature Importance

- *SHAP Value*

In our feature importance analysis, SHAP is also a well-known method in post-model analysis to compare and analyze the final features, since it generates numeric values that can be used to calculate the important role of the features to the model. They illustrate the positive or negative impact of each feature on the prediction for that instance. The top 20 features of Shap value are shown in the following figure.

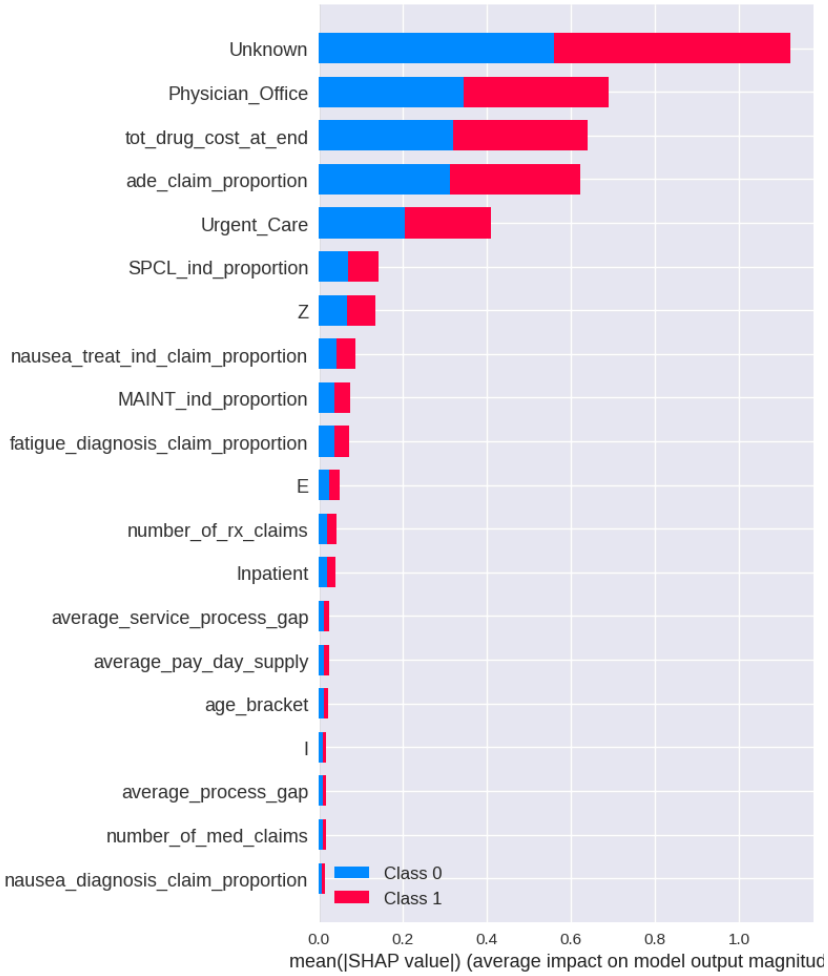


Figure 9: SHAP Values

Place-of-Treatment Related Factors: Unknown, Physcian_Office, Urgent_Care, are features representing the total counts of each patient's place of getting treatment. This indicates that the treatment location of patients is strongly correlated with osimertinib discontinuation. Specifically, the count of physician office visits claims has the most

predictive power in the LightGBM model. We found Patients with fewer physician office visits are more likely to discontinue the therapy.

Side-effect Related Factors:
ade_claim_proportion, fatigue_diagnoisis_claim_propotion, nausea_treat_ind_claim_proportion, as well as nausea_diagnosis_claim_propotion represents the ratio of how often a specific side effect is reported in their medical claim history. A higher proportion indicates that the side effect has been reported frequently in their medical claim history. These features indicate that the various side effects patients are experiencing have predictive power in the model.

Financial-related Factors: tot_drug_cost_at_end represents the financial cost of medication for patients.MAINT_ind_proportion and SPCL_ind_proportion are the proportions of patient claims for these medications across all their pharmacy claims. As we know, maintenance drugs are usually used for chronic diseases, which might have higher cumulative costs. Meanwhile, specialty drugs are usually more expensive than non-specialty drugs. These financial-related factors show high predictive power in the model. We also found that patients with lower spending amounts are more likely to discontinue therapy.

Disease-Related Factors: We have extracted disease letter codes from the original ICD-10 codes. Each ICD-10 code starts with a specific letter, indicating the classification of the associated disease. For instance, the code "L" stands for skin-related issues and demonstrates predictive power in our model.

5.2 Relationship Between Factors

To further analyze the important features and the relationship between these factors, we used SHAP dependence plots to study the individual effects and interaction effects of key variables. We will skip the feature 'Unknown' at this stage.

- **ade_claim_proportion & physician office**

The following two dependency plots illustrate the relationship between 'physician_office' and 'ade_claim_proportion.' In the first plot, It is evident that, for a given 'ade_claim_proportion,' patients who visit 'physician_office' less frequently are more likely to be predicted as unsuccessful (tgt_ade_dc_ind == 1).

To further investigate this, we examined the interaction effect of 'physician_office' with 'fatigue_diagnosis_claim,' the most common adverse event found in the dataset. We observed that after around 10 visits to the physician's office, the proportion of 'fatigue_diagnosis_claim' starts to decrease.

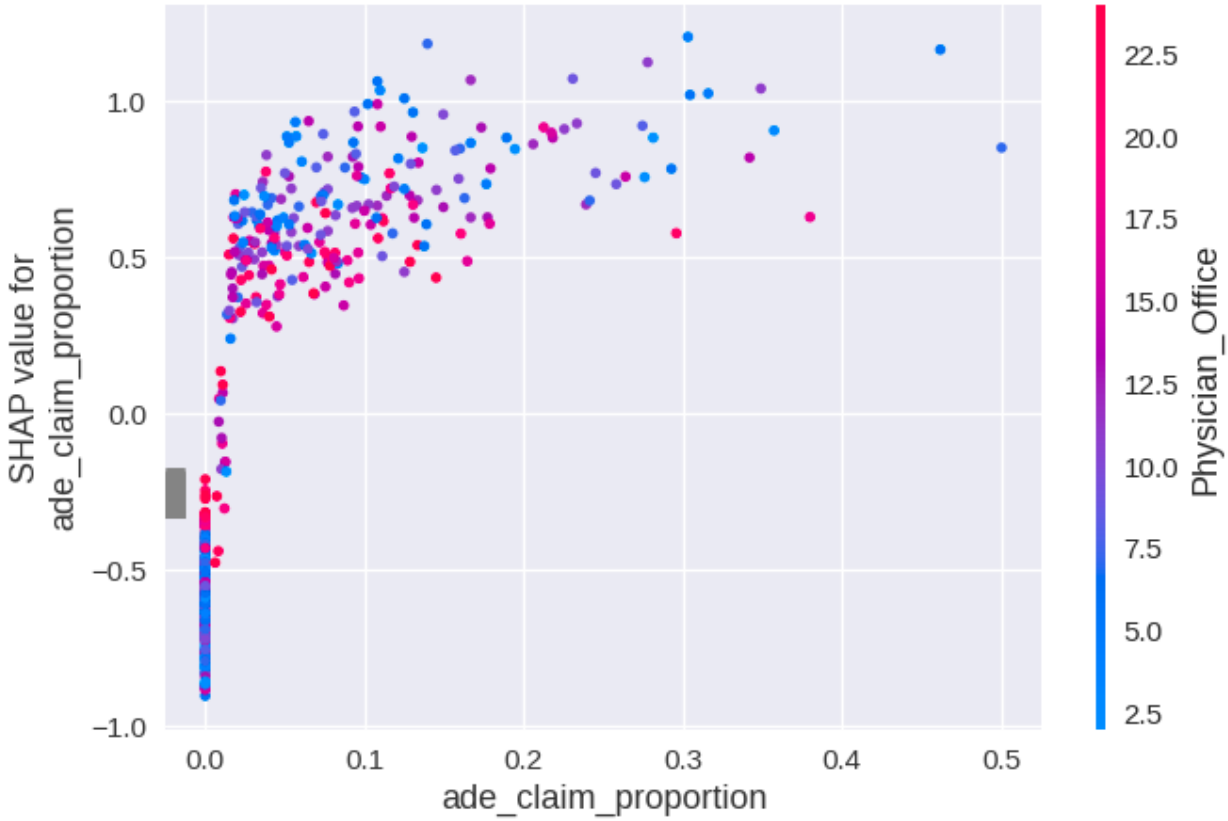


Figure 10: SHAP dependence For ADE Claim Proportion and Physician Office

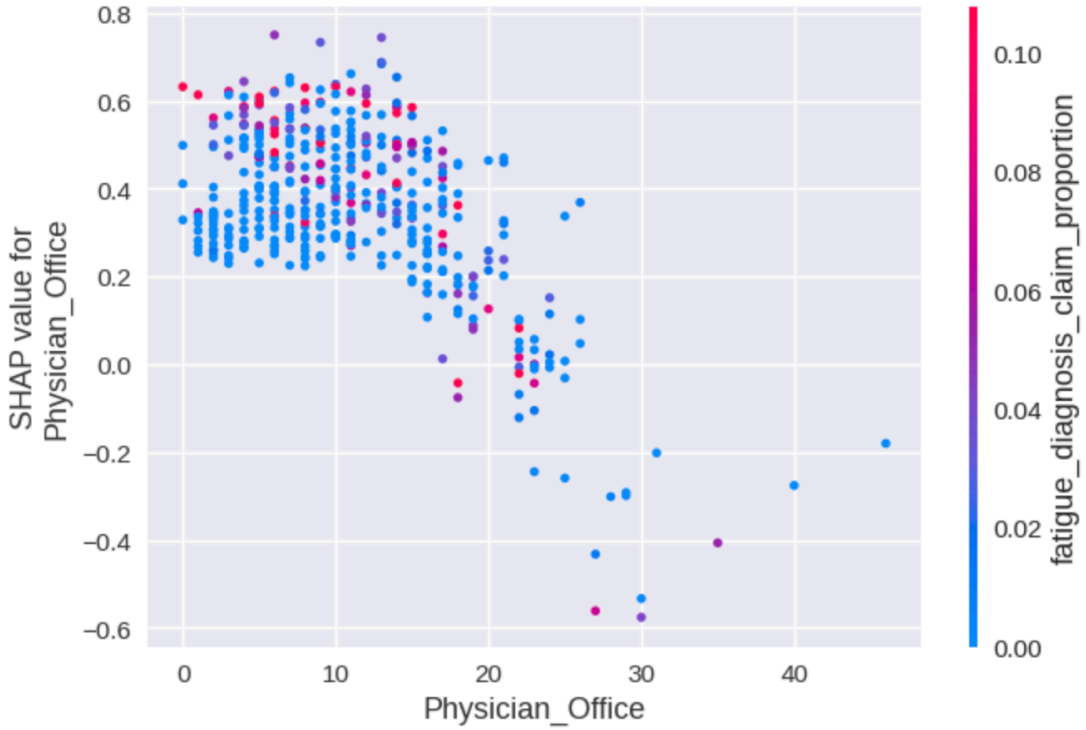


Figure 11: SHAP dependence For Physician Office and Fatigue_diagnosis_claim_proportion

- **ade_claim_proportion & tot_drug_cost_at_end**

The dependency plot below illustrates the relationship between 'ade_claim_proportion' and 'tot_drug_cost_at_end.' It is evident that, for a given 'ade_claim_proportion,' patients who spend more during the therapy duration are more likely to be predicted as unsuccessful.

Upon further examination of the dataset, we found that individuals who spend more are also more likely to order specialty and branded medications, which are more expensive than generic medications. Additionally, they tend to have a higher number of medical and pharmacy claims. We will provide information about where their spending is allocated in the upcoming segmentation section.

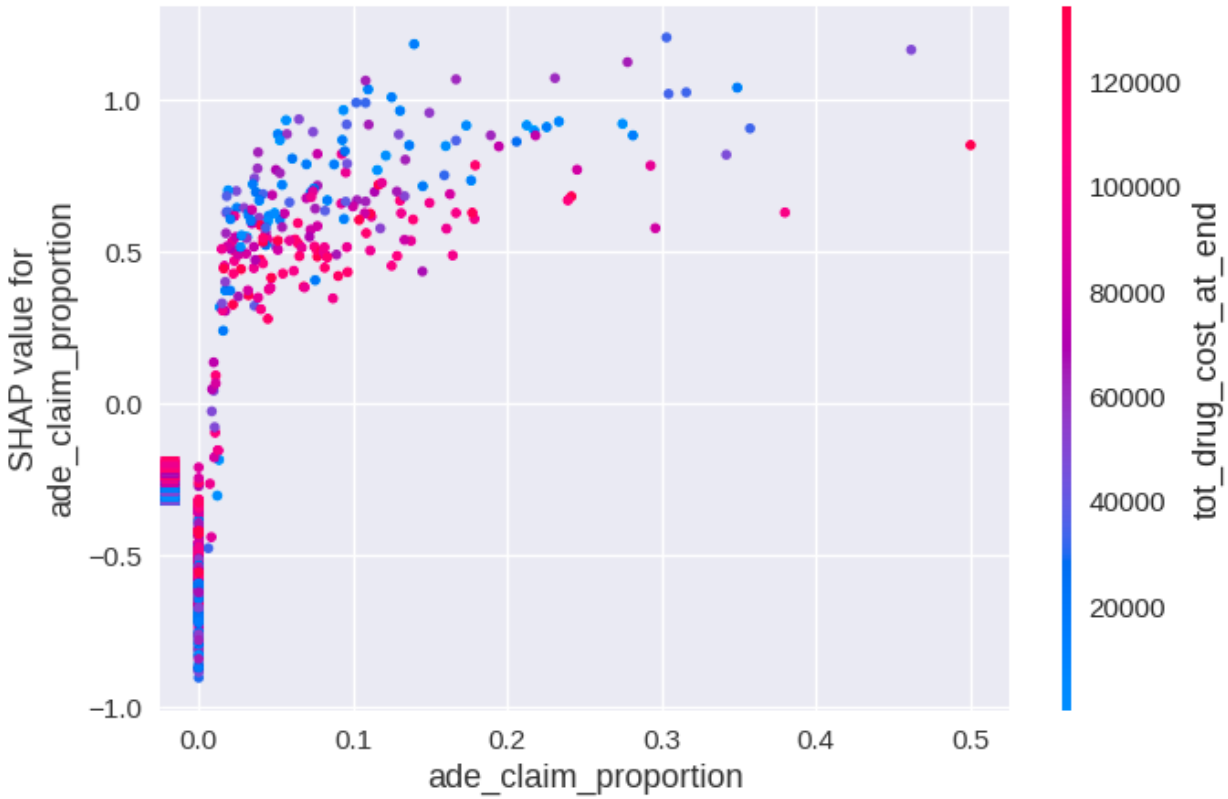


Figure 12: SHAP dependence Values For ADE Claim Proportion and Total Drug Spending

6. SEGMENTATION

6.1 Segment Features

Based on our research and feature analysis using our predictive model, it has become evident that the issue of osimertinib non-adherence can be attributed to various factors among our members. To gain a deeper understanding of these underlying causes and effectively devise tailored recommendations for improving osimertinib adherence, we need to categorize our members into distinct segments.

According to the importance and relationship analysis of features we generated before, we selected the following 3 key features in different fields and separated all members into 3 segments:

- *Physician_Office:* Represents the number of times the patient visited a physician's office.
- *tot_drug_cost_at_end:* Reflects the total amount the patient spent during the therapy.
- *ade_claim_proportion:* Indicates the proportion of patients who reported adverse events among all their medical visits.

We applied the K-means clustering method to divide the patient population into four distinct segments. However, given that over 50% of patients lack medical claim records, we faced uncertainty regarding their ADE reporting and treatment locations. Consequently, we focused on segmenting the 523 members who had medical records related to treatment locations and ade_claim_proportion.

Within each segment, we computed the average tgt_ade_dc_ind, indicating the level of treatment success for that segment. Concurrently, we also calculated tot_drug_cost_at_end, ade_claim_proportion, number_of_med_claims, cms_disabled_ind, and cms_low_income_ind to gain insights into potential factors contributing to non-adherence across different segments.

The summarized conditions and values of each segment are presented in the table below.

Cluster Variable	Segment 1	Segment 2	Segment 3
Physician_Office	8.59	13.97	13.8
tot_drug_cost_at_end	24,350	1,37,053	85,811
ade_claim_proportion	0.05491	0.04657	0.05173
Group size			
Number	219	76	228
Percentage	41.87%	14.53%	43.59%
Other Factors			
number_of_med_claims	53	67	73
cms_low_income_ind	34.25%	59.21%	37.72%
cms_disabled_ind	15.53%	30.26%	19.74%
Target variable			
tgt_ade_dc_ind	32.88%	5.26%	14.91%

Figure 13: Characteristics of Different Segments

The figure above describes the result for each segment, and each group is classified using Physician_Office,tot_drug_cost_at_end, and ade_claim_proportion variable. The average of tgt_ade_dc_ind is different from each segment and they are also different in number_of_med_claims ,cms_low_income_ind and cms_disabled_ind. Therefore, the segmentation of all members is reasonable and meaningful to help better understand the reason behind Osimertinib non-adherence problem amongst members.

6.2 Segments Analysis

Segment 1: Non-Low-Income Patients with Inactive Physician Office Visits

This segment represents individuals not identified as low-income by CMS and exhibit inactivity in physician_office visits. This group accounts for 41.87% of the members. Among them, 32.88% have discontinued treatment and reported adverse drug events, the highest percentage compared to other segments. Compared with segment 3, which also consists of non-low-income patients, this group spends an average of $6,000 less, indicating their cost-consciousness. This suggests that those who are very cost-conscious and display an inactive pattern in physician_office visits are more likely to experience osimertinib non-adherence issues.

To understand why this segment is most prone to osimertinib non-adherence problems, we conducted secondary research.

We found a meta-analysis on Physician Communication and Patient Adherence to Treatment that showed a significant positive correlation between physician communication and patient adherence. Patients with poor physician communication are 19% more likely to be nonadherent compared to those with effective communication. This aligns with our modeling, emphasizing 'physician_office' as a strong predictor of treatment success. Effective physician communication involves providing constant support, genuine empathy, and adequate relief for patient distress. Studies on lung cancer patients have shown higher distress levels compared to other cancer patients. Physician office visits not only offer physical support like pain management but also emotional support. Regular follow-up visits are crucial for monitoring treatment progress, managing side effects, and providing emotional support.

Segment 2: Low-Income Patients with Active Physician Office Visits

This segment represents individuals identified as low-income by CMS and exhibiting activity in physician_office visits, making up 14.53% of the members. Among them, only 5.26% have discontinued treatment and reported adverse drug events, which is the lowest percentage among all segments. This group demonstrates non-cost-conscious behavior, as they tend to spend significantly more than the other two segments, possibly due to government subsidies. It suggests that individuals who are not cost-conscious and maintain an active pattern of physician_office visits are less likely to face osimertinib non-adherence issues. Additionally, it's interesting to note that this group has the highest disability rate, indicating that disability is a common trait among low-income individuals.

Segment 3: Non-Low-Income Patients with Active Physician Office Visits

This segment comprises individuals not identified as low-income by CMS and exhibit activity in physician_office visits, making up 43.59% of the members. Among them, 14.91% have discontinued treatment and reported adverse drug events, which is the second highest compared to other segments. This group also displays moderate cost-

conscious behavior, spending a moderate amount during treatment. Compared to non-low-income patients in Segment 1, they have fewer financial problems; therefore, other factors may be contributing to their treatment discontinuation.

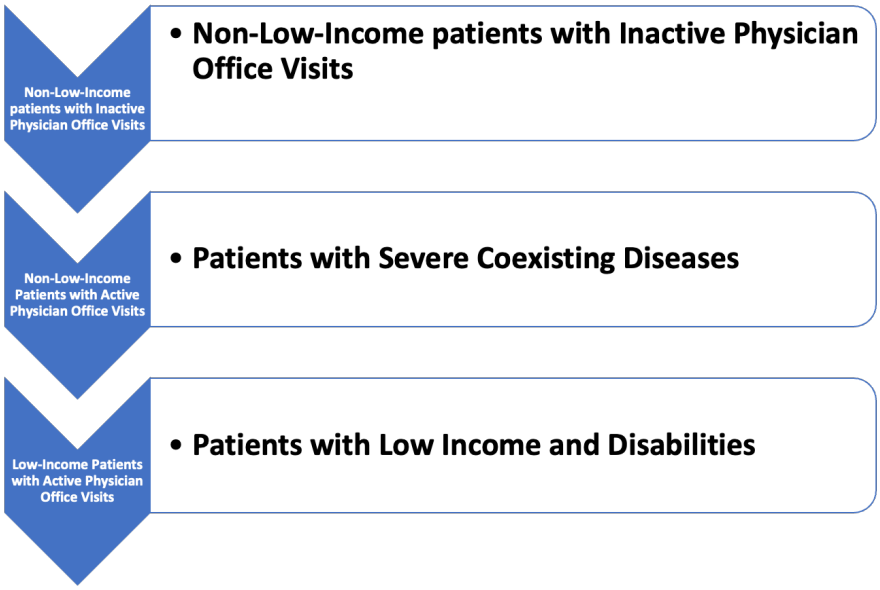
Further investigation into their diagnosis history revealed that Segment 3 had a higher prevalence of diagnoses classified under "A," representing infectious and parasitic diseases. This was primarily due to a high number of primary diagnosis claims related to A419, corresponding to salmonella sepsis. The most common symptom of this condition is a recurring fever. Additionally, salmonella sepsis could be an acute condition that requires immediate medical attention. This suggests that, in addition to adverse drug events related to osimertinib, patients are consistently experiencing discomfort due to other co-existing diseases.

Relationships of Segments and Diseases

Variable	Segment 1	Segment 2	Segment 3
Diagnosis Letter Code: A	0.2511	0.1053	0.4474

7. RECOMMENDATIONS

Based on our segmentation, we have identified three main clusters of members based on economic factors, patients' health status, and adherence status. Since members facing medication adherence issues are not homogeneous and may have different reasons for their medication continuation, we need to identify these reasons and provide tailored solutions for each segment. To enhance medication adherence for a broader range of individuals, we ranked our member clusters using urgency and feasibility metrics. Subsequently, we provided specific solutions for each target segment, aiming to address the needs of as many members from diverse segments as possible.



Our priority is to address the financial and low physician office visit challenges faced by patients who are not low-income and have an inactive pattern of physician office visits. They are the most likely group to experience osimertinib non-adherence, which is a concern far beyond that of other groups. Secondly, we are focusing on patients who have severe coexisting diseases, such as salmonella sepsis, causing them significant suffering. Unlike other patients, these individuals experience fewer financial challenges and maintain an active pattern of physician office visits. The presence of coexisting diseases results in persistent pain and discomfort, making them less tolerant of osimertinib. Following that, our attention will be on the low-income group with disabilities. Although these individuals actively visit physicians, their disabilities require us to provide them with more convenient access to the medication and care they need. Finally, we will present a strategy that we have found to be beneficial for all members in improving their osimertinib adherence.

7.1 Strategy Program

Everyone's cancer journey is different, so it's important to provide treatment plans that are tailored to each individual's needs and goals.

7.1.1 Segment 1: Providing Financial Assitance and Home Visits

We understand Cancer treatment can be expensive, and no one should have to choose between paying for their medication and paying for other essential expenses. As we analyzed above, the reason this segment is the most likely to suffer from osimertinib non-adherence is their cost-conscious nature; Patients may not buy the next batch of medicines or may not visit the physician due to financial constraints. According to the analysis, patients who have lower physician office visits are more likely to discontinue the treatment. Our mission is to alleviate financial concerns and provide personalized monitoring that is lacking due to low physician office visits.

Implementation:

1. **Extending Payment Duration:** Provide extended payment durations to reduce their monthly financial constraints. As such, rather than making upfront payments, patients can be allowed to make the payment in equated monthly installments (EMI). For the same, patients could be charged a nominal interest rate without hurting their pockets and Humana's business.

2. **Nurse or Healthcare Professional Compulsory Home Visits**: Collaborate with DispatchHealth for bi-weekly in-home Nurse Visits: Assign nurses or healthcare professionals to schedule up to 2 compulsory monthly visits to the patients' homes, not just to monitor the patient's health and administer treatments, but most importantly to foster relation with the patient, provide education on mitigating the side effects with proper diagnosis, the importance of the Osimertinib medication and treatment, availability of financial assistance programs etc, resulting in increased medical adherence.

7.1.2 Segment 2: Providing Transportation Services and Educating Patients about Mail Delivery Service

Getting to and doctor's appointments and other healthcare facilities can be difficult for people with cancer, especially those who live in rural areas or who don't have access to a car. Based on the segmentation analysis conducted earlier, low-income patients identified by CMS display active physician visit patterns. However, this group of people also exhibits a high rate of disability, signaling challenges in commuting to the physician's office. It is crucial to provide convenient transportation support and efficient mail delivery services to ensure that they can consistently attend their medical appointments and access the medication they need.

Implementation:

1. **Partner with NGOs, Nonprofit or Community-based Organisations:** The organizations can assist by providing their workers who can assist in transporting patients to their medical appointments. This will be a cost-effective solution for the patients, resulting in an increase in the count of patients adhering to their appointments.

2. **Educate patients about online prescription refills.** Humana needs to implement a program to inform patients about the availability of the mail-delivery service and ensure they have access to this service. This can be achieved by educating them on how to use online prescription refills especially for ADE or phone-based refill options.

7.1.3 Segment 3: Providing Tailored Medication Treatment and Monitoring

As we have previously analyzed, patients in Segment 3 exhibit fewer financial problems and already maintain an active pattern of physician visits. The presence of coexisting diseases could be the primary reason for their non-adherence to osimertinib. For patients with severe coexisting diseases, it is crucial to provide tailored medication treatment and appropriate monitoring, as these coexisting diseases can have interaction effects that reduce the effectiveness of osimertinib. Therefore, we recommend that Humana offers customized medication treatment based on the specific coexisting diseases each patient has and monitors the other medications they are taking.

Implementation：

1. **Create an 'Easy Report' function on the website or app,** allowing for the tracking of patients with severe coexisting diseases, documenting their entire treatment journey, therapy experiences, and monitoring progress. Based on the monitoring data, personalized feedback and recommendations can be provided to patients with lung cancer who are taking osimertinib, thereby offering them tailored medication treatment.

2. **Establish connections between patients and their designated physicians**.It is important for patients to connect with their designated physicians especially in

the case of reported side effects, as the physician usually holds an in-depth understanding of their coexisting diseases and can provide personalized medication treatments.

7.1.4 Segment 4 (All Members): Providing Medication Administration Reminder and Monitoring.

As we mentioned in the data exploration stage, we noticed that many patients tended to discontinue their treatment precisely at the end of the first 30 days, with subsequent peaks in discontinuation occurring at 60 and 90 days. This suggests that patients typically complete one package of medication before discontinuing treatment. Consequently, it is crucial to monitor patients' medication administration history and provide timely reminders to encourage them to refill their medication before running out. Additionally, we found that patients who visit their physician's office more than 10 times have a lower incidence of ADE claims. This suggests that patients who visit their doctor more frequently are more likely to have fewer ADEs. Therefore, it is advisable for patients to maintain regular contact with their physician, especially if they have complex medical conditions.

Implementation:

1. **Provide Medication Reminders to Patients**: Once the first medicine batch of Osimertinib is completed for the patient, Humana can give a *courtesy call* to the patient to check on the patient's health status and if the patient needs assistance to reorder or refill the medicine. Typical check-ins with patients between days 25 and 29 to remind them of their upcoming refill could be helpful. Humana can implement reminders via push notifications through apps or emails when patients are about to complete their monthly cycle of osimertinib medication.

2. **Subscription Model:** Humana can offer a subscription model where customers commit to receiving regular 60-day supplies over an extended period. This model can also provide up to 2 Physician appointments every month at a discounted rate(5-10%). This subscription model can not only eliminate the burden of re-fills by automatically delivering medications to patients' homes regularly, but also improves medication adherence through personalized medical support

- Offer bundled medicine packages with EMI options (Equated Monthly Installment): This can help patients to reduce the burden of huge upfront

payment on their 60-day medications. By offering nominal discounts for purchasing medications in bulk, Humana could encourage patients to choose 60-day supplies instead of the common 30-day purchase, motivating them to continue medication for more than 30 days.

Note: Discount here means that Humana bears 5-10% of the amount the patient originally owes in their co-pay.

3. **Incentivise Monthly Scans Coupled With Physician Office Visits:** To encourage more people to continue their medications for more than 30 days, patients could be recommended to mandatorily undergo a scan (like a CT scan) in 45 days (15 days before the next peak dropout (60days) to assess their health status, followed by a physician's office visit or virtual healthcare appointment (especially for the elderly) to discuss the scan results and next steps. Seeing objective evidence of recovery and receiving support from a healthcare provider can ensure patients are taking their medications as prescribed, that any side effects are being monitored and encouraged to continue the therapy.

4. **Customized App Features for Osimertinib:**

   Humana can incorporate features tailored to osimertinib therapy, encompassing the tracking of the entire patient journey, monitoring ADE symptoms, providing a Q&A section, and FAQs specifically for managing common ADE symptoms such as fatigue and nausea (most frequently observed in patients who discontinue the therapy). Additionally, educational materials can be provided to inform patients about the therapy and reassure them about the normalcy of side effects with correct diagnosis.

7.2 Cost & Effectiveness Analysis

After we segmented members, we quantified the cost of our recommendations, and analyzed if the total cost of recommendations can offset or be lower than the current cost to Humana (total insurance claim fee) by implementing the above data-driven and actionable recommendations.

1. Cost Saving (Save more, Earn more Model):

If higher proportion of patients complete the treatment, they'll not only have 80% less chance of recurrence of the cancer but will also not have to undergo advanced-level lung cancer treatments like chemotherapy, radiation therapy or other surgeries,

resulting in cost saving for Humana. Further, we have provided below the cost Humana will save as the proportion of successful treatment increases.

The total cost of the patients who are suffering from lung cancer but haven't successfully completed the treatment will be large as these patients are prone to suffering from advanced-level treatment (therapy, surgery etc) in the future.

C1: Cost Post-Unsuccessful Treatment

As the disease advances, the need for surgery, therapy and specialized care increases, leading to sky-rocketting costs**.** *Therefore*, the C1 category would have a total estimated cost of $2,00,000 per patient.[1]

C2: Cost of Successful Treatment

The average cost range of a patient who successfully completes the treatment from the data provided is *$80,000 per patient.*

(S1-S4 pertains to each segments mentioned above in the recommendation)

S1: Providing Financial Assistance and Home Visits

Firstly, we are suggesting Humana to provide Segment 1 with financial assistance to extend their payment duration and reduce their monthly cost. The total amount charged from the patients would remain the same, so there is no additional cost. Regarding the nurse home visits, which are already a service provided by Humana, we suggest optimizing the frequency of these visits. Our findings indicate that after 10 visits to the in-person assistance provided at the physician's office, the reported ADE begins to decrease. *Therefore, there won't be any additional cost for Category S1.*

S2: Providing Transportation Services and Educating Patients about Mail Delivery Service

The cost of providing transportation Services will include reimbursement charges for the driver. On average, reimbursing volunteers for mileage, parking, and other expenses related to driving patients to their appointments would likely be around $15-$20 per trip, amounting to approximately $40 per month per patient. We need to educate and train disabled patients about the online and phone-based refill service, which might primarily occur through Humana's internal team. The additional cost for this education is expected to be low.

Therefore, the cost for category S2 would be *$40 per month per patient. As the entire treatment timeline is 6 months (or 180 days), the total cost for this category will be $40*6 = $240 per patient.*

S3: Providing Tailored Medication Treatment and Monitoring

Amongst the sugested in-app features, including ‘Easy Report’ function could be added to Humana's app by their technical team, likely at a lower cost.

In long term, integrating needed features, specially customized for members with lung cancer, into the existing app. If Humana opts to build the feature on their app, the upfront cost is around $5,000 [2]; however, as benefits last beyond 6 months, the monthly cost allocation will be low. *Therefore, it would be at a nominal cost of $5000-$7000*

S4: Providing Medication Administration Reminders and Monitoring.

Adding a In-App Push Notification Feature: This simple feature will be integrated into Humana's app by their technical team, likely at a lower cost. The subscription model will necessitate a 5-10% discount from Humana, aligning with Humana's individual maximum out-of-pocket range of $3,500–$7,350 [3], translating to about $452 per month per person.

While the scans we suggested are by default required, we propose scheduling these scans around 45 days to allow the effects of osimertinib to manifest, and before 60 days, which marks the next peak of patients discontinuing the therapy, to ensure the reports are reviewed by the patients. This is likely to significantly improve their adherence rate. The cost for this suggestion would be $0. *Therefore, the cost for category S4 would be around $450 per month per person. As the entire treatment timeline is 6 months (or 180 days), the total cost for this category will be $450*6 = $2700 per patient.*

Total Cost Saved by Humana as unit-increase in successful treatment count = (C2 + S1 + S2 + S3 + S4) - C1 = ***$1,14,060 per patient*.**

Through our analysis and data-driven recommendations, Humana will not just be able to *save more, earn more* but also be able to have a *successful treatment rate, save more lives* and *build a higher reputation*.

2. Brand Reputation: Through implementing these recommendations, as the successful treatment proportion increases, Humana will be able to save more lives as the medication effectiveness leads to 80% less chance of lung-cancer

recurrence and twice as likely to survive from the disease. Our recommendations will not just assist Humana to save more lives, but also build a bigger and reliable brand image and reputation, resulting in more popularity and will also lead to more people joining the Humana insurance services.

8. Future Scope

1. Through preliminary and granular analysis, we saw that the two datasets (pharmacy claim and medical claim) do not have a primary key to link the two datasets in a way that every patient's medical claim information is accurately linked with the type of medicine the patient bought for that particular diagnosis. This can happen if there's a primary key unique for every patient and every diagnosis of the patient.

2. When using feature importance to identify key modeling indicators, the top-ranked feature pertains to the 'unknown' category of treatment location. Since 'unknown' exhibits strong predictive power in determining whether patients are likely to discontinue therapy before 180 days and report adverse drug events, we recommend that Humana improves the data collection process to uncover the actual treatment locations of these patients. Doing so would provide significant benefits in identifying target patients and gaining valuable business insights.

3. When analyzing common factors for lung cancer and factors affecting therapy adherence, It would be helpful to have data on the patient's smoking behavior. This is because smoking has been one of the most common reasons for lung cancer. We could have also predicted better if therapy adherence is lower for those with smoking behavior/history. In addition, members' data on the type of body scans, and frequency would have helped us understand if there existed any correlation between regular body scan checkup counts and therapy adherence.

9. Conclusion

In predicting members who are most likely to experience Osimertinib discontinuation and report adverse drug events, we first analyzed the dataset and prepared the data for modeling. Subsequently, we applied logistic regression, XGBoost, and LightGBM for preliminary prediction and compared their performances along with the corresponding AUC values. According to our model, LightGBM has the best performance with an AUC of 0.958 for the training dataset, outperforming the other models. In the segmentation

analysis based on physician office visit frequency, the amount spent, and ADE claim proportion, we utilized the K-means clustering method to segment members into three groups. For each of these groups, we put forward targeted and personalized recommendations, including providing financial assistance, coordinated home care visits, transportation support, regular push notifications for patients to access their medication, and necessary monitoring during the therapy. With these measures, Humana can improve Osimertinib adherence and seek the greatest benefit for its members.

Through our suggestions, Humana can help save lives and make a real difference in the world. Humana can create a place where everyone has access to the healthcare they need, and the support they need to stay on track with their treatment and achieve the best possible health outcomes.

By showing compassion and understanding, we hope to enable Humana to help its members through the difficult journey of small-cell lung cancer treatment. People with cancer need to know that they are not alone and that there are people who care about them and want to help them. Our suggestions aim to let Humana make a real difference in the lives of its members and create a more humanitarian world.

Humana can embody this spirit by helping its members through the difficult journey of cancer treatment.

"The greatest good is what we do for one another"- Mother Teresa

10. References

[1] Doe, J. (2022, July 12). The Cost of Lung Cancer Treatment. Verywell Health. https://www.verywellhealth.com/lung-cancer-treatment-cost-5217734

[2] Vaniukov, S. (2022, August 19). How Much Does It Cost to Develop an App in 2023. Softermii. https://www.softermii.com/blog/how-much-does-it-cost-to-make-an-app

[3] Humana. (n.d.). Copay Health Insurance for Large Group and Small Business. Retrieved from https://www.humana.com/employer/products-services/medical-plans/copay-plans

[4] American Cancer Society. (n.d.). Lung Cancer Statistics | How Common is Lung Cancer? Retrieved from https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html

[5] Lynch J, Goodhart F, Saunders Y, O'Connor SJ. Screening for psychological distress in patients with lung cancer: results of a clinical audit evaluating the use of the patient Distress Thermometer. Support Care Cancer. 2010 Feb;19(2):193-202. doi: 10.1007/s00520-009-0799-8. Epub 2010 Jan 13. PMID: 20069436; PMCID: PMC3016098.