OPTIMIZING ONCOLOGY:

A DATA-DRIVEN APPROACH TO ENHANCE OSIMERTINIB THERAPY ADHERENCE

2023 HUMANA-MAYS
HEALTHCARE ANALYTICS
CASE COMPETITION

Executive Summary

This analysis aims to identify patients at risk of prematurely discontinuing Tagrisso lung cancer treatment due to adverse drug events (ADEs). We utilized the medical and pharmacy claim data provided by Humana on patients before and during their treatment from the years 2018-2022. The data was preprocessed to engineer relevant features like diagnosis counts, hospital visits, and costs, which could be fed into a machine-learning model. Several machine learning models were developed and evaluated, with XGBoost selected as the top performer due to its high AUC-ROC and generalizability to new and unseen data. The model was made more accurate and robust with hyperparameter tuning.

Analysis of the feature importances of the model highlighted factors like the timing of the first ADE, age, and circulatory disorders as the most influential. A SHAP analysis was conducted to quantify each feature's impact on predictions. Key insights were translated into targeted interventions focused on patient education, accessibility, financial assistance, and side effect mitigation for high-risk groups. Concrete ROI projections demonstrated positive cost-benefit outcomes.

In summary, this project combined ethical data practices, rigorous methodology, and translational understanding to generate actionable insights that can significantly improve Tagrisso adherence. Our recommendations take into account different subgroups as well as their particular needs in order to save Humana at least 20 million dollars a year – even if only one of them is utilized. Overall, the project provides a valuable framework for leveraging analytics to tackle complex healthcare issues and improve patient outcomes.

Table of Contents

Case Background

Within the realm of oncology, Tagrisso has emerged as a beacon of hope for patients combating early-stage lung cancer of the non-small cell variety, distinguished by a specific targetable mutation known as EGFR. Despite its proven efficacy, the side effects associated with Tagrisso, from nausea and fatigue to pain and high blood glucose, often lead patients to discontinue their therapy prematurely.

Humana, a leading health insurance company, is deeply committed to patient welfare and uses advanced data analytics to elevate patient experiences and treatment efficacy. Through in-depth analysis of patient interactions with Tagrisso, they found that within the first six months of this life-saving medication, approximately one-quarter of Humana members succumbed to these adverse drug events (ADEs), derailing their treatment course.

Predicting premature therapy discontinuations due to adverse drug events (ADEs) is crucial, showcasing Humana's dedication to leveraging data-driven insights for continuous, life-saving patient care.

Problem Statement

Against this backdrop, Humana faces a crucial mission: address the high rate of premature discontinuation of Tagrisso therapy due to ADEs. The organization has a vast dataset from 2018 to 2022, including detailed insurance claims and a 90-day individual lookback period.

The objective is clear — identify members who face the risk of discontinuing their Tagrisso therapy due to ADEs and intervene proactively. By delving into this dataset, the challenge is to not only predict premature therapy discontinuation but also understand the intricate web of factors leading to this outcome.

The crux of the issue lies in predicting which patients are at risk of premature therapy termination due to these ADEs, a challenge exacerbated by the complex interplay of individual health profiles, therapy adherence patterns, and specific drug interactions. Success in this endeavor means more than just accurate predictions; it signifies the ability to offer timely interventions, ensuring that patients vulnerable to ADE-driven therapy discontinuation receive the support they need. The challenge, therefore, lies in unraveling the intricate patterns within the data, leveraging the historical context, and employing advanced analytics to empower healthcare providers with the foresight to intervene effectively and keep the flame of hope alive for those battling against cancer.

Description of Data Provided

The given training data is presented as three separate CSV files: target data, medical claims, and pharmacy claims. In total, this represents a training dataset of 1232 unique patients with 100,159 medical claims and 32,133 pharmacy claims. Amongst these records, there are 57 distinct feature columns, and each column's meaning is described in Humana's provided data dictionary, which is not posted here for brevity.

Structure of Report

We will first outline the steps we took to engineer insightful features from the data, how we dealt with missing values, and give an exploratory analysis of the dataset. Then, we will outline which machine learning models we considered, how we decided on the optimal one, and the steps we took to improve the model fit. We will then discuss the feature's importance based on our final model. Afterward, we will leverage the insights we gained to give concrete recommendations to Humana on how to approach preventing people from dropping out of treatment due to adverse side effects while establishing key performance indicators to keep track of.

Data Analysis

Overall Methodology

The three datasets provided for each of the train and holdout sets did not lend themselves to easy analysis. We needed somehow to combine the datasets such that there was a single entry for each person, with relevant feature columns. Therefore, our first step was to engineer new features, doing some cleaning in the meantime. We created numerous features from the medical and pharmacy claim datasets to capture as much of the relevant information as possible. Then we performed the final data cleaning and did an exploratory data analysis before feeding them into an ML model.

Feature Engineering

There is not a 1:1 relationship between the three given datasets. For example, one patient in the target dataset may have zero medical claims, one medical claim, or multiple medical claims; another patient may have >= 0 pharmacy claims.

Therefore, we required features that represent aggregations of medical claims and pharmacy claims for each individual patient. We chose aggregations that would capture the most context behind the associated features for each individual patient. Our feature

creations were motivated by linking features with the psychological/physiological reasons that would plausibly lead to a patient having an ADE diagnosis and quitting their treatment. We generated a total of 74 features, which we grouped into categories for ease of explanation. Below is a description of the features in each category, and how missing data was handled, if applicable.

The first step in our data engineering process was to combine the 3 holdout and 3 train files into single datasets. This allowed us to more easily generate features and identify relationships across the data. The medical and pharmaceutical claims were merged into a single data frame with NaN values for the opposite claim type. This enabled grouping by patient while still identifying claim types. The identity data (e.g., sex, race) was then joined to add demographic info to each patient's claims. These combined datasets were created for both training and holdout to enable consistent feature engineering.

ADE Diagnoses

Whether someone has been diagnosed with an ADE is obviously very relevant for our predictions. So, our first feature was the number of ADE diagnoses recorded in each patient's medical claims. We also added a feature for the number of days between the start of treatment and the first ADE diagnosis (first_ade_date) and the last ADE diagnosis (last_ade_date), along with the mean day of all ADE diagnoses. We predicted that these would sufficiently capture the ADE diagnosis portion of our response variable and would be a critical indicator of whether a patient would quit treatment.

Table 1: ADE Diagnoses Features

Feature Name	Description	Data Type
n_ade_diagnoses	Number of ADE diagnoses per patient	Integer
first_ade_date	Days between start of treatment and first ADE	Integer
last_ade_date	Days between start of treatment and last ADE	Integer
mean_ade_day	Average day of ADE	Integer

If someone had no associated medical claims, their n_ade_diagnoses were set to -1, and the first, mean, and last date columns were set to -91, 181, and 181, respectively. This was done to help the model distinguish them from other patients who may have had continuous medical claims without any ADEs.

Hospital Visits

To capture hospital visit data, we first created an indicator column, any_medical_claim, to specify if the patient had any associated medical claims. We grouped medical claims by the place of treatment and calculated the number of hospital visits associated with visits to the ER, ambulance visits, and acute care. Those without any visits had these filled as 0.

We calculated how many days after starting Tagrisso the patients made their first and last hospital visits, filling those without any visits as 181. This could serve as an indication of when the patients first started experiencing problems after taking Tagrisso and whether they stopped going to the hospital prematurely. We used the visit_date field in the original medical claim dataset to calculate our day features. Finally, we added the total number of hospital visits and of individual medical claims by counting the medclm_key and clm_unique_key columns in the original dataset, respectively.

Table 2: Hospital Visits Features

Feature Name	Description	Data Type
any_medical_claim	Whether patient has made any medical claims	Boolean
ambulance_visits	Number of total ambulance visits by patient	Integer
er_visits	Number of total ER visits by patient	Integer
acute_visits	Number of total medical non-ER,non-ambulance visits by patient	Integer
n_med_claims	Number of medical claims	Integer
n_med_visits	Number of medical visits	Integer
first_med_claim_date	Days from start of treatment to first medical claim	Integer
last_med_claim_date	Days from start of treatment to last medical claim	Integer

Major Diagnoses

The original medical claims data helpfully includes separate columns for well-known side effects of Tagrisso. We calculated the total number of all such diagnoses for each

patient, filling patients without medical claims as 0. We would expect more side effect diagnoses to play a major role in patients' cessation of treatment, especially in the absence of usage of relevant therapeutic drugs, which we also captured in our data with relevant features, as will be described shortly.

Table 3: Major Diagnoses Features

Feature Name	Description	Data Type
n_seizure_diagnoses, n_pain_diagnoses, n_fatigue_diagnoses, n_nausea_diagnoses, n_hyperglycemia_diagnoses, n_constipation_diagnoses, n_diarrhea_diagnoses	Number of all such diagnoses	Integer

Minor Diagnoses

Each medical claim has up to 8 non-primary diagnoses listed in ICD-10 format. We downloaded a list of such codes and associated diagnoses online and used this to count the number of all such non-primary diagnoses **[1]**.

The presence of separate diagnoses can often be a symptom of treatments and may influence an individual to drop their treatment. Those without medical claims had these filled as 0.

Table 4: Minor Diagnoses Features

Feature Names	Description	Data Type
Infectious Diseases, Neoplasms, Blood Disorders and Immune Disorders, Endocrine and Metabolic Diseases, Mental and Neurodevelopmental Disorders, Nervous System Disorders, Eye and Ear Disorders, Circulatory System Diseases, Respiratory System Diseases, Digestive System Diseases, Skin and Subcutaneous Tissue Diseases, Musculoskeletal and Connective Tissue Diseases, Genitourinary System Diseases, Pregnancy and Perinatal Conditions, Congenital Malformations and Abnormalities Symptoms and Signs, Injuries and External Causes, Special Codes, Health Status and Contact with Health Services	Number of medical claim ICD-10 codes corresponding to each disease/disorder	Integer

Diagnoses Totals

We created several aggregate features to capture some more insights from the major and minor diagnoses data, such as the maximum number of unique simultaneous diagnoses and the mean number of diagnoses per visit, for both minor and major diagnoses, as well as their combined sum. These are described below.

Table 5: Diagnoses Totals Features

Feature Name	Description	Data Type
max_simulatenous_diagnoses	Maximum number of simultaneous distinct medical diagnoses	Integer
mean_simulatenous_diagnoses	Average number of simultaneous diagnoses	Integer
max_major_diagnoses	Maximum number of simultaneous major (Tagrisso ADE-related) diagnoses	Integer
mean_major_diagnoses	Average number of simultaneous major (Tagrisso ADE-related) diagnoses	Integer
max_minor_diagnoses	Maximum number of simultaneous minor (non-Tagrisso ADE-related) diagnoses	Integer
mean_minor_diagnoses	Average number of simultaneous minor (non-Tagrisso ADE-related) diagnoses	Integer

Pharmacy Claims

The pharmacy claim data can provide crucial insights as to whether patients are treating their medical conditions. We added features similar to those for medical claims. Additionally, we counted the number of mail orders patients had before and after treatment, which may help flag those who are unable to go to the pharmacy and therefore may be in a more precarious situation. Those without pharmacy data had the first and last dates filled with -91 and 181 respectively, and the other features as 0. We used the service_date field in the original pharmacy claim dataset to calculate our day features.

Table 6: Pharmacy Claims Features

Feature Name	Description	Data Type
any_pharmacy_claim	Whether patient has made any pharmacy claims	Boolean
first_pharmacy_claim_date	Days from start of treatment to first pharmaceutical claim	Integer
last_pharmacy_claim_date	Days from start of treatment to last pharmaceutical claim	Integer
mail_orders_before	Number of mail ordered drugs for patient before treatment	Integer
mail_orders_after	Number of mail ordered drugs for patient after treatment	Integer
n_pharmacy_claims	Number of pharmacy claims	Integer

Drugs

The pharmacy claims data specifies if the claim was for a drug with a known interaction with Tagrisso (ddi_ind) and whether it is used to prevent the common Tagrisso side effects we described in the Major Diagnoses section. Hence, we counted the total number of Tagrisso ADE symptom-related drugs and drugs with ddi_ind =1 for each patient. Another feature for the number of unique drugs taken by a patient, calculated by counting the unique ndc_ids in the pharmacy claims, was added for each patient.

We also created a column for the number of unique drugs a patient took for their treatment of chronic illnesses. We hypothesize that people with chronic conditions may be more vulnerable to burnout in treatment. Those without any pharmacy claims had these filled as 0.

Table 7: Drugs Features

Feature Name	Description	Data Type
unique_maint_drugs	Number of unique drugs patient has taken for chronic conditions	Integer
unique_non_maint_drugs	Number of unique drugs patient has taken for non-chronic conditions	Integer
n_unique_drugs	Number of unique drugs taken by patient	Integer
ddi_drug_count	Number of unique drugs taken for drugs with direct interactions with Tagrisso	Integer
anticoag_drug_count	Number of unique drugs taken for anticoagulant drugs	Integer
diarrhea_drug_count	Number of unique drugs taken for diarrhea drugs	Integer
nausea_drug_count	Number of unique drugs taken for nausea-related drugs	Integer
seizure_drug_count	Number of unique drugs taken for seizure-related drugs	Integer

Pharmacy Costs

Any factors prohibiting easy access to treatment drugs were also considered, as drug costs can play a major role in patient withdrawal from treatments. We added total pharmaceutical costs pre and post-treatment to capture this. We also noted the maximum and average cost of drugs per patient to get features that added more context to typical and outlier purchases. We retained the tot_drug_cost_accum_amt feature present in the original data, which describes the cumulative cost amount for a member year-to-date, as it sometimes seemed to differ from our calculated totals for the year and might, therefore, have some additional data in it.

Finally, we added the number of unique specialty drugs patients used, as these tend to be especially expensive drugs. These data could especially be useful when combined with the indicator for whether a patient is classified as low-income. Those without pharmacy claims had these features filled as 0.

Table 8: Pharmacy Costs Features

Feature Name	Description	Data Type
rx_cost_before	Total pharmaceutical cost for patient before treatment start	Integer
rx_cost_after	Total pharmaceutical cost for patient after treatment start	Integer
rx_cost_total	Total pharmaceutical cost for patient over all known data	Integer
rx_cost_ytd_totals_acc	Total year-to-date cost of pharmaceuticals for patient	Integer
max_drug_cost	Maximum cost of a single drug for patient	Integer
avg_drug_cost	Average cost of all drugs for patient	Integer
specialty_drug_count	Number of unique drugs taken in specialty drugs	Integer

Personal Attributes

Finally, we kept the personal attributes in the original dataset to see if our model would even make use of them. If so, we could take steps to deal with inequitable predictions, especially if the model considered race and sex as predictors.

We also added a cluster feature that we got from K-means clustering on the dataset. K-means clustering is a partitioning method that divides a dataset into K distinct, non-overlapping subsets (clusters) based on the distances between data points. Incorporating cluster labels as features can capture some inherent structures or patterns in the data, potentially enhancing the performance of supervised classifiers by providing an additional layer of context or grouping.

Table 9: Personal Attribute Features

Feature Name	Description	Data Type
race_cd	Race of patient	Integer
est_age	Age of patient	Float
sex_cd	Sex of patient	Boolean
cms_disabled_ind	Whether a patient is classified as disabled	Boolean
cms_low_income_ind	Whether a patient is classified as low-income	Boolean
cluster	Assigned cluster from K-means clustering	Integer

Principal Component Analysis

In order to determine which of our features should be utilized by the final model, we performed a principal component analysis on the 74 features that we engineered. PCA would be very helpful in either dropping some features beforehand or realizing which features should ideally be avoided in our models. The principal component analysis revealed several key insights about the sources of variance in our dataset.

The top principal components were able to explain a substantial portion of the total variance - the first component explained 22.8%, and the top 5 components together explained around 50%. This suggests there are some inherent dimensions of dimensionality reduction within the data that can be captured in fewer components. Examining the loadings of the individual features provided further detail about which elements were driving the components. Features relating to overall medical utilization and timing, such as the number of medical claims, visits, and cost totals, generally had the highest loadings. This could indicate that a patient's overall use of medical services is a primary source of variance.

Demographic features like race, sex, and age had very low loadings in comparison, meaning they explain little of the total variance - the data does not appear heavily dependent on these attributes, which is instrumental in making our conclusions and models as bias-free as possible.

Some specific clinical features stood out as moderately important in driving the components. Diagnoses related to symptoms, neoplasms, and circulatory system diseases had higher loadings, implying some of the variance relates to these underlying conditions. Pharmacy features like drug counts and costs were also moderately loaded, suggesting medication usage patterns also relate to patient differences.

In summary, the PCA indicates that utilization, timing, diagnoses, and medications are key sources of variance, while demographics are less important. The data has inherent dimensionality that can be reduced while preserving the most important information. Instead of instantly removing the "less useful" features we decided to instead create models and let the model decide which features it wants to use as some features might be more or less important depending on the model. Our conclusions from PCA would instead be utilized in making sure the model did use the most important features and wasn't heavily indexing into the less important ones. This way we could maintain high levels of accuracy while still avoiding overfitting on a few numbers of variables.

Further Data Cleaning

Variable Encoding and Integer Conversion

In data preprocessing, we first dropped the therapy_start_date column since it was not needed for modeling. The sex_cd column was encoded as 0 for female and 1 for male to convert categories into numeric values. Boolean columns like any_medical_claim and any_pharmacy_claim were also converted to integers for consistency. Missing values were filled in through random sampling based on column proportions (for race_cd, sex_cd, cms_disabled_ind, cms_low_income_ind) or mean imputation (for age). This step removed all missing values from the data.

Erroneous Data Points

13 people in the train set were found to have a label 1 for discontinuing treatment, yet had no ADE Diagnosis reported. This could be explained by the relevant medical claim data being missing. Or the label data was incorrect since the project prompt states that an ADE has to be reported for the person to count as discontinuing treatment due to ADE. To be on the safe side, we removed these 13 people from the dataset. The final XGBoost model also unsurprisingly does not label anyone without an ADE diagnosis as 1.

Additionally, some patients seemed to have claim data that extended beyond the -90 to 180-day period that the project prompt stated we should have. We clipped the date values outside the 90-day lookback window to be -90 or 180 days.

Matched Distributions for Equity

To enable matched distributions between train and holdout, we performed clustering on the combined dataset and separated the clusters back out. Before feeding the data into the final model, the overrepresented clusters in train data were resampled to better

match the cluster proportions in holdout data. This created a more consistent distribution between the two datasets by only duplicating a few data points.

As a final step, the target and ID columns were added back to the preprocessed train and holdout dataframes. Throughout our data preprocessing, we were cognizant of ensuring unbiased, equitable treatment of all individuals in the datasets. By filling in missing demographic values like race, sex, and income status through random sampling, we retained the original distribution of these protected attributes rather than introducing skew. Our clustering and downsampling steps aligned the train and holdout distributions without considering any sensitive subgroups - this prevented imbalances that could disproportionately impact certain populations. Dates and clinical values were cleaned based on domain knowledge, not any specific attributes of the patients. Overall, our data cleaning methodology focused on formatting and completeness of the data while consciously avoiding the introduction of bias. By promoting balance and fairness at this initial step, we established a solid framework for ethical and equitable machine learning modeling moving forward.

Exploratory Data Analysis

After creating our numerous features and combining all the train and holdout data to their respective tables we conducted an EDA analysis to get a glimpse into the distribution of certain features and the target label, to see if we needed to take special action to achieve equitable outcomes or rebalance our datasets.

Attribute and ADE Tables

Below are tables describing how ADE diagnoses and therapy discontinuation are distributed by patients' personal attributes for the training data. Note that These tables have been generated after the 13 erroneous entries have been removed. Overall, the tables below showed no inherent bias in the data with regard to personal attributes.

Table 10: Race and ADE

Race	Number Reporting ADEs	Number Discontinuing Treatment	Total Number
0	11	6	151
1	206	67	692
2	56	13	144
3	11	4	36
4	20	6	149
5	10	7	42
6	1	1	5

Table 11: Sex and ADE

Sex	Number Reporting ADEs	Number Discontinuing Treatment	Total Number
0	241	78	865
1	74	26	354

Table 12: Disability and ADE

Disability	Number Reporting ADEs	Number Discontinuing Treatment	Total Number
0	257	89	1038
1	58	15	181

Table 13: Low income and ADE

Low Income	Number Reporting ADEs	Number Discontinuing Treatment	Total Number
0	189	66	751
1	126	38	468

Table 14: Race and Sex by ADE

Race	Sex	Reports Ade	Discontinues	Total Count
0	0	8	6	101
0	1	3	0	50
1	0	157	50	496
1	1	49	17	196
2	0	47	9	110
2	1	9	4	34
3	0	8	3	24
3	1	3	1	12
4	0	12	4	102
4	1	8	2	47
5	0	9	6	30
5	1	1	1	12
6	0	0	0	2
6	1	1	1	3

Table 15: Missing Data in Train and Test Sets

Data	Train		Test	
	# Missing	% Missing	# Missing	% Missing
Medical claims	696	57.1%	235	56.0%
Pharmacy claims	71	5.8%	41	9.8%
Race	68	5.6%	18	4.3%
Sex	83	6.8%	28	6.7%
Low income indicator	83	6.8%	28	6.7%
Disability Indicator	83	6.8%	28	6.7%
Sex	83	6.8%	28	6.7%

Correlation Matrices

Correlation matrices were generated for each of the buckets of features, as generating a single matrix for all 70+ features would produce unwieldy results. Correlation Matrices can be used to get a sense of how strongly each feature may affect the target variable and in what directions. However, it's not possible to tell how features may work together to influence the results. A couple of the most interesting ones are shown below.

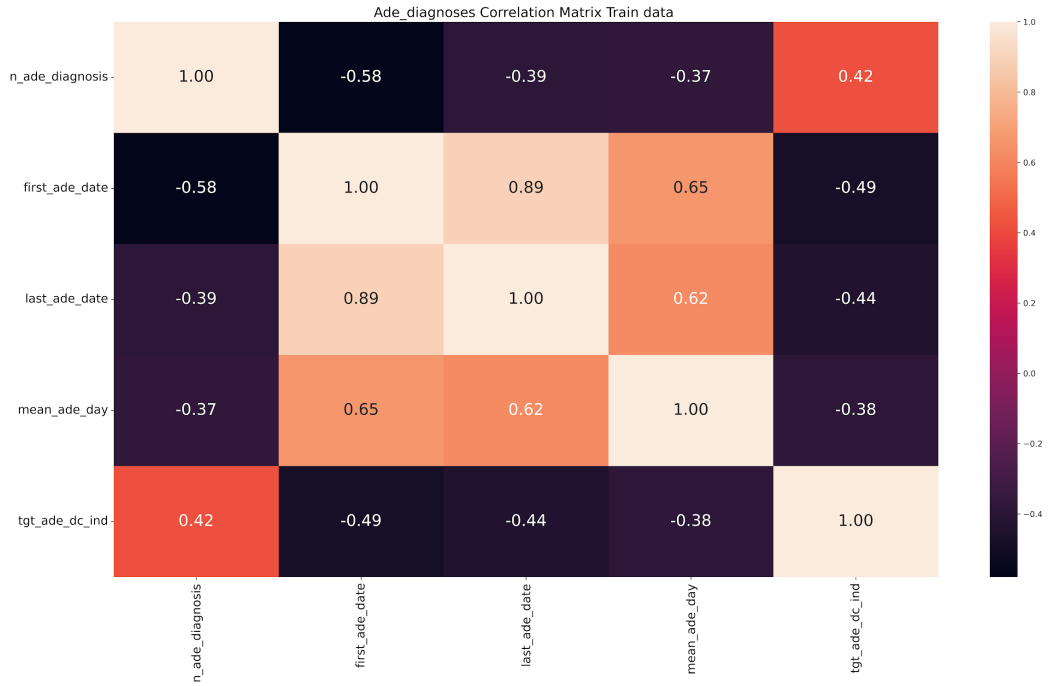

Figure 1: Ade Diagnoses Features Correlation Matrix

The above matrix shows, naturally, that higher numbers of n_ade_diagnoses are positively correlated with the target discontinuing therapy (see rightmost column). On the other hand, the three date features suggest that the earlier theADEis diagnosed, the less likely the patient will discontinue.

The last column of Figure 2 below shows that the target is not correlated with personal attributes at all, which should help our model achieve equitable outcomes.

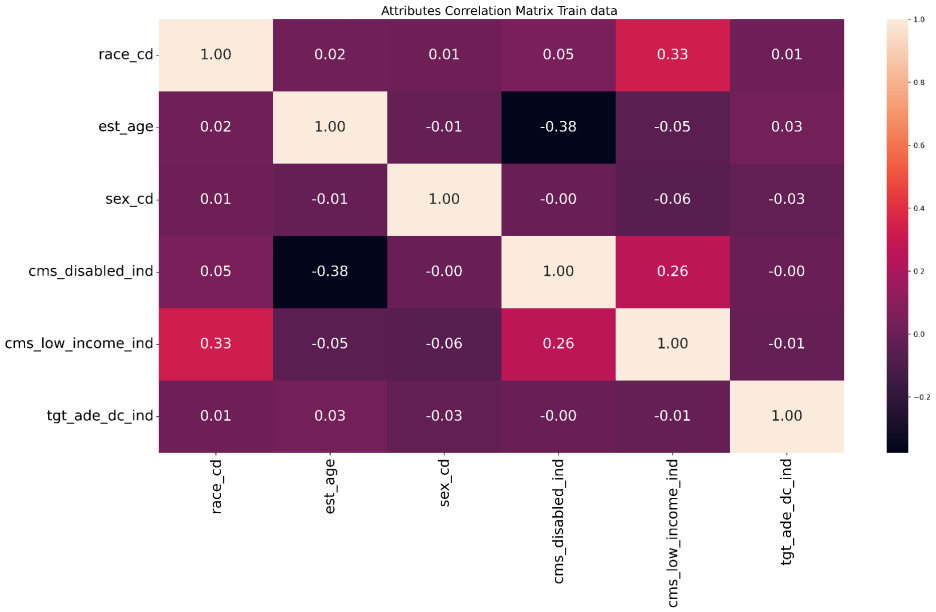

Figure 2: Attribute Features Correlation Matrix

Finally, the matrix below shows that metrics that represent sum total and total simultaneous diagnoses all correlate positively with our target. In contrast, the features representing the number of drugs being taken have negligible correlations (matrix too big to show here), suggesting that people being diagnosed and not taking drugs could be more likely to discontinue treatment.

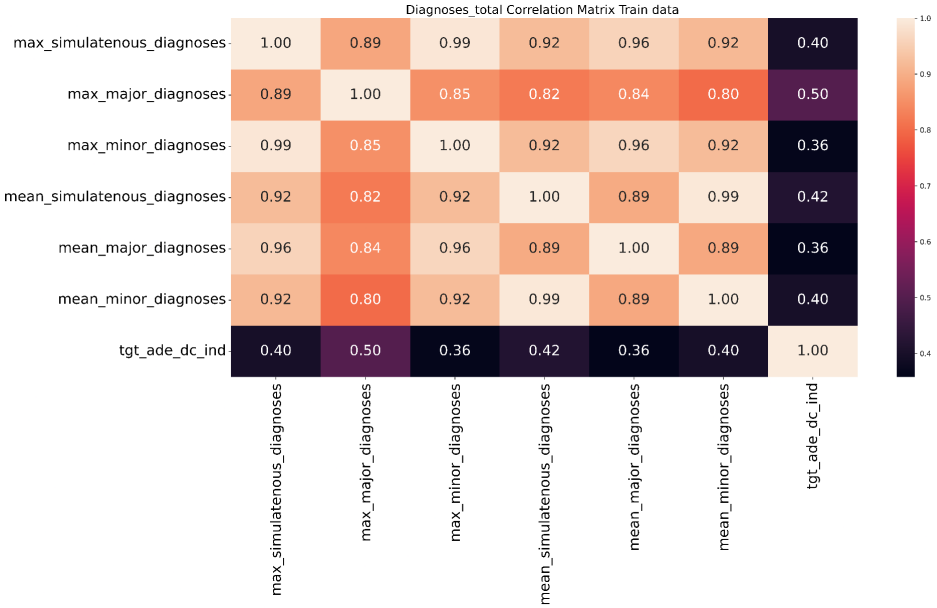

Figure 3: Diagnoses Totals Features Correlation Matrix

Train vs Holdout Data

Distribution of Personal Attributes

Table 16: Distribution of Personal Attributes in Train and Test Tables

Field	Value	Train %	Test %
Race	0	12.4%	11.4%
	1	56.8%	52.1%
	2	11.8%	14.5%
	3	3.0%	5.7%
	4	12.2%	11.9%
	5	3.4%	4.3%
	6	0.4%	0.0%
Sex	0	71.0%	70.0%
	1	29.0%	30.0%
Disability indicator	0	85.2%	86.7%
	1	14.8%	13.3%
Low-income indicator	0	61.6%	62.4%
	1	38.4%	37.6%
Age	Mean	73.7%	73.1%

From the table above we observe that there is no significant difference between the train and test dataset with regards to patient distribution. The table below shows that there are some discrepancies in the missing data counts, most importantly in pharmacy claim data. But on the whole are numerous features for both medical and pharmacy claims means almost every patient should have enough associated features, excepting for a few which are discussed in a late section.

Table 17: Missing Data Counts in Train and Test Tables

Data	Train		Test	
	# Missing	**% Missing**	**# Missing**	**% Missing**
Medical claims	696	57.1%	235	56.0%
Pharmacy claims	71	5.8%	41	9.8%
Race	68	5.6%	18	4.3%
Sex	83	6.8%	28	6.7%
Low-income indicator	83	6.8%	28	6.7%
Disability Indicator	83	6.8%	28	6.7%
Sex	83	6.8%	28	6.7%

Distribution of Feature Values

Histograms were generated to compare the distribution of features across the train and holdout sets for all variables. They were normalized to indicate the probabilities of each respective bin, and a continuous probability distribution line was fitted. On the whole inspection of histograms did not highlight any obvious discrepancies between the train and holdout patients. The histograms were either fairly uninformative (as in Figure 4) or showed differences (as in Figure 5) that could be explained by the number of missing data. For instance, the reason the test set patients appear to use fewer unique maintenance drugs in Figure 5 is due to the fact that a higher percentage of the test set (9.8% vs 5.8% in the train) lacks pharmacy data, and therefore is said to use 0 maintenance drugs.

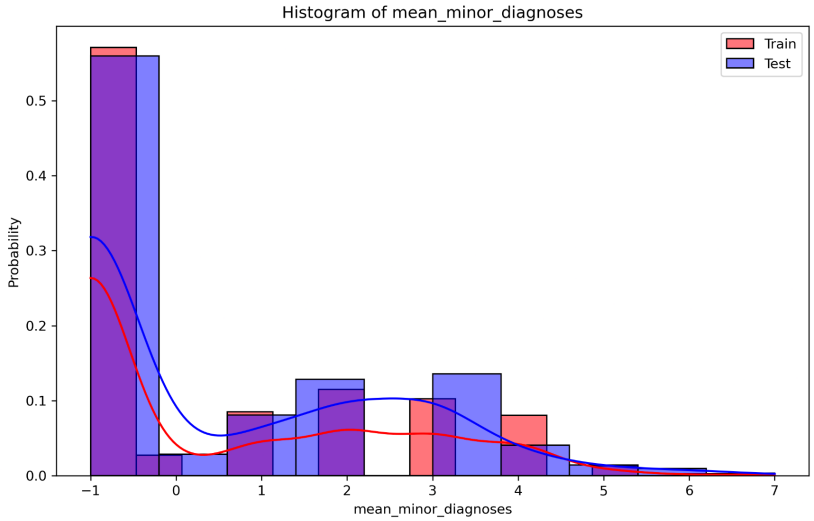

Figure 4: Histogram of average number of minor diagnoses

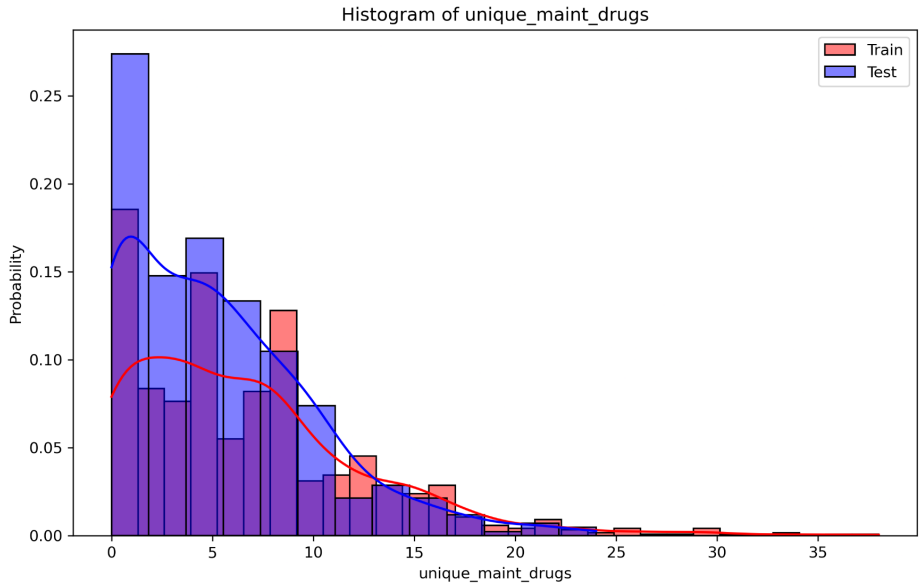

Figure 5: Histogram of Number of Unique Maintenance Drugs

Missing Data Uncertainties

It is unclear whether the many people who are missing medical claim and pharmacy claim data either made no such claims or are simply missing this data. Our approach to overcome this uncertainty was to include indicator columns to specify whether someone had no pharmacy or medical claim data, and to fill the features related to such claims with sensible filler values, usually of -1, 0 for drug and diagnoses data, and -91 or 181 for date related data. In any case, those without medical data are never classified as discontinuing treatment since they have no ades reported.

Those who could pose real difficulty for the model are patients without pharmacy data, yet with medical data that shows at least one ADE diagnosis. Luckily, as shown in Table 14 below, there are very few such people. Our model may have difficulty making accurate predictions for these people without significant pharmacy-related features, but there are still lots of medical claim related features available for them.

Table 18: At-risk patients missing substantial data

Type	# in Train	# in Test
Patients with some medical data but no pharmacy data	12	10
Those above who have an ADE reported	9	5
Those above who discontinue treatment	6	NA

Modeling

Model Objective

The fundamental objective of our predictive model is to identify patients at risk of prematurely discontinuing their Tagrisso therapy due to adverse drug events (ADEs). In the context of imbalanced data and the critical nature of healthcare interventions, the model must excel at both sensitivity and specificity. Sensitivity is key because it ensures the identification of all actual positive cases (ADE-induced discontinuations) among those at risk, thus minimizing false negatives and enabling timely interventions. Specificity, on the other hand, ensures that the model correctly identifies patients who do not face ADE-induced discontinuations, preventing unnecessary interventions and conserving resources.

Simply put, the model aims to predict if a patient will stop using Tagrisso within 180 days while also minimizing bias to ensure fair treatment decisions for all patients.

Key Performance Indicators - Model

The key performance indicators for our model will be ROC-AUC score, and other accuracy metrics such as accuracy, sensitivity, and specificity. Sensitivity measures the proportion of true positives correctly identified, while specificity measures the proportion of true negatives correctly identified. It is crucial our model performs well on both counts to both correctly identify those patients most at risk of dropping out, while minimizing false positives so Humana does not waste resources on preventative measures on people who are not at risk of dropping out.

Additionally the model has to be generalisable, meaning it performs well on new and unseen data. Improving these metrics will directly translate to better identification of

at-risk patients, more timely interventions, and cost savings from avoided hospitalizations.

Model Approach & Selection

We considered three machine learning models, each with different strengths, that were commonly used for binary classification tasks and which each team member had experience working with.

Logistic Regression

Logistic Regression is a statistical method for modeling binary outcomes. It estimates the probability that a given instance belongs to a particular category based on one or more independent variables by fitting the data to a logistic curve. We choose this foundational model for its simplicity and interpretability, offering a benchmark against which more complex models could be evaluated.

Random Forest

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees for a given input. We chose it for its ensemble learning and ability to handle non-linear relationships, which could provide robustness against overfitting while allowing it to capture complex interactions within the data.

XGBoost

XGBoost, or eXtreme Gradient Boosting, is an optimization of gradient boosting machine learning that constructs a model in a stage-wise fashion like other boosting methods. However, it specifically adjusts model performance with the goal of reducing prediction error.

Hyperparameter Tuning

We implemented Grid Search to find the optimal hyperparameters for the above 3 models based on the 10-fold cross-validation score. In 10-fold cross-validation, the data is divided into 10 equal parts. One part is used as validation, while the other nine are as training. This process is repeated 10 times, each time using a different part for validation. The results are averaged for a single performance estimate, indicating the model's generalization ability.

Table 19: Grid Search Results for three models

Model	Search Grid	Best Params	10-Fold CV Score
Logistic	C': [0.01, 0.1, 1, 10, 100], 'penalty': ['l1', 'l2', 'elasticnet'], 'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']	C': 100' penalty': 'l2' 'solver': 'newton-cg'	0.721
Random	n_estimators': [50, 100, 150], 'max_depth': [10, 20, 30, 40], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4], 'bootstrap': [True, False]	"bootstrap': False 'max_depth': 10 'min_samples_leaf': 4 'min_samples_split': 2 'n_estimators': 50	0.902
XGBoost	max_depth': [3,4], 'learning_rate': [0.01, 0.05, 0.1], 'n_estimators': [50, 100, 150], 'reg_alpha': [0, 0.05, 0.1], # L1 regularization term 'reg_lambda': [0, 0.05, 0.1] # L2 regularization term	learning_rate = 0.05 max_depth = 4 n_estimators = 100 reg_alpha = 0.0 reg_lambda = 0.1	0.912

Model Results

To analyze each model's performance further, we split our train set 85%-15% to train and validation sets, using stratified sampling for race and sex to make sure the datasets weren't imbalanced. The final results of these are given below:

Model Performances

The following table highlights each model's performance.

Table 20: Model Confusion Matrices

Model	Confusion Matrix - Train Set		Confusion Matrix - Validation Set	
Logistic Regression	0.98	0.02	0.96	0.04
	0.27	0.73	0.33	0.67
Random Forest	1.00	0.00	0.99	0.01
	0.01	0.99	0.24	0.76
XGBoost	1.00	0.00	1.00	0.00
	0.08	0.92	0.21	0.79

Table 21: Confusion matrix Format

		Prediction	
		Positive	Negative
Actual	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Table 22: Model Performance Metrics

Model	Accuracy		Recall		Specificity		Precision	
	Train	Val	Train	Val	Train	Val	Train	Val
Logistic Regression	0.86	0.82	0.78	0.74	0.97	0.94	0.98	0.96
Random Forest	1.00	0.88	0.99	0.74	1.00	0.94	NA	0.96
XGBoost	0.96	0.90	0.93	0.83	0.96	0.95	0.99	0.96

Why XGBoost

From the tables above it is apparent that XGBoost consistently outperformed the other models in almost all metrics. Logistic Regression was too simple a model to capture the inherent complexities in the, while Random forest ended up overfitting the training set. XGBoost emerged as the frontrunner due to its ability to handle imbalanced datasets and learn intricate patterns. Its ensemble of decision trees and regularization techniques mitigates overfitting, ensuring reliable generalization to unseen data. Additionally, XGBoost's innate capacity to handle missing values, a common challenge in real-world datasets, enhanced its suitability for our task.

Thus XGBoost best met our key performance indicators as it achieved a high ROC-AUC score in the 10-Fold Cross Validation, demonstrating its robustness against overfitting. It does this while striking a balance between sensitivity and specificity. Thus it not only accurately identifies patients at risk but also minimizes unnecessary interventions with false negatives, highlighting its effectiveness in a real-world healthcare context.

XGBoost Advanced HyperParameter Tuning

We did some further hyperparameter tuning to get the best possible XGBoost model. We expanded our grid search parameter space significantly and also ran a Random Search. Unlike grid search, random search samples hyperparameter combinations from predefined distributions randomly rather than exhaustively. It can be faster and

sometimes more effective than grid search. We employed Randomized Search to explore a broad hyperparameter space efficiently.

Our hyperparameter tuning approach (a blend of RandomizedSearchCV and GridSearchCV) was tailored to our problem's intricacies, ensuring our XGBoost model was finely tuned to navigate the complexities of the dataset. The table below shows our search parameters and the resulting optimal parameters and 10-fold validation scores.

Table 23: Hyperparameter Tuning for XGBoost

Search Type	Search Grid	Best Params	10-Fold CV Score
Random Search (with 10000 iterations)	max_depth': randint(2, 6), 'learning_rate': uniform(0.03, 0.5), 'n_estimators': randint(70, 150), 'reg_alpha': uniform(0, 0.1), 'reg_lambda': uniform(0.01, 0.2)	learning_rate': 0.11068367919419156 'max_depth': 3 'n_estimators': 90 'reg_alpha': 0.08054323292999864 reg_lambda': 0.16203218595803182	0.9958
Grid Search	'max_depth': [2,3,4,5], 'learning_rate': [0.03, 0.04, 0.05, 0.06, 0.07, 0.08], 'n_estimators': [80, 90, 100, 110, 120, 130, 140], 'reg_alpha': [0, 0.01, 0.02], 'reg_lambda': [0.09, 0.1, 0.11]	'max_depth': 4 'learning_rate': 0.04 'n_estimators': 115 'reg_alpha': 0.01 'reg_lambda': 0.1	0.9278

The randomized exploration allowed us to cover diverse configurations, ensuring a comprehensive search for optimal combinations. However, the optimal parameters yielded an unseemingly high ROC-AUC score of 0.9958. This meant the search actually overfitted our parameters, as a slight change in the parameters drastically affected our score.

The Grid Search results were more promising. The optimal parameters gave a commendable ROC-AUC score of 0.9278. This score did not change noticeably for slight changes in parameters, suggesting a good, generalizable fit.

The chosen parameters reflect a delicate equilibrium, emphasizing both accuracy and generalizability, thus ensuring the model's reliability in real-world applications:

- **Learning Rate:** A balanced learning rate (0.04) ensures that the model converges to an optimal solution without overshooting or getting stuck in local minima.

- **Tree Depth (max_depth):** A moderate tree depth (4) strikes a balance between capturing intricate patterns and preventing overfitting.
- **Number of Estimators (n_estimators):** 115 estimators provide a robust ensemble, capturing diverse patterns within the data while averting overfitting.
- **Regularization Terms (reg_alpha and reg_lambda):** Minimal regularization (0.01 and 0.1 respectively) prevents excessive complexity, maintaining model simplicity without compromising predictive power.

Final XGboost Model Training

We then fit our model to the training data with one final nuance: Early Stopping. Early Stopping is a regularization method where the training process is halted as soon as the performance on a validation set starts to degrade, as this indicates potential overfitting, rather than continuing until the training data is perfectly fitted. The figure below shows the roc score during the training of our model on the train and validation set, with the horizontal dashed line indicating where we ceased training. The epoch corresponds to the “number of estimators” parameter in the XGBoost tree.

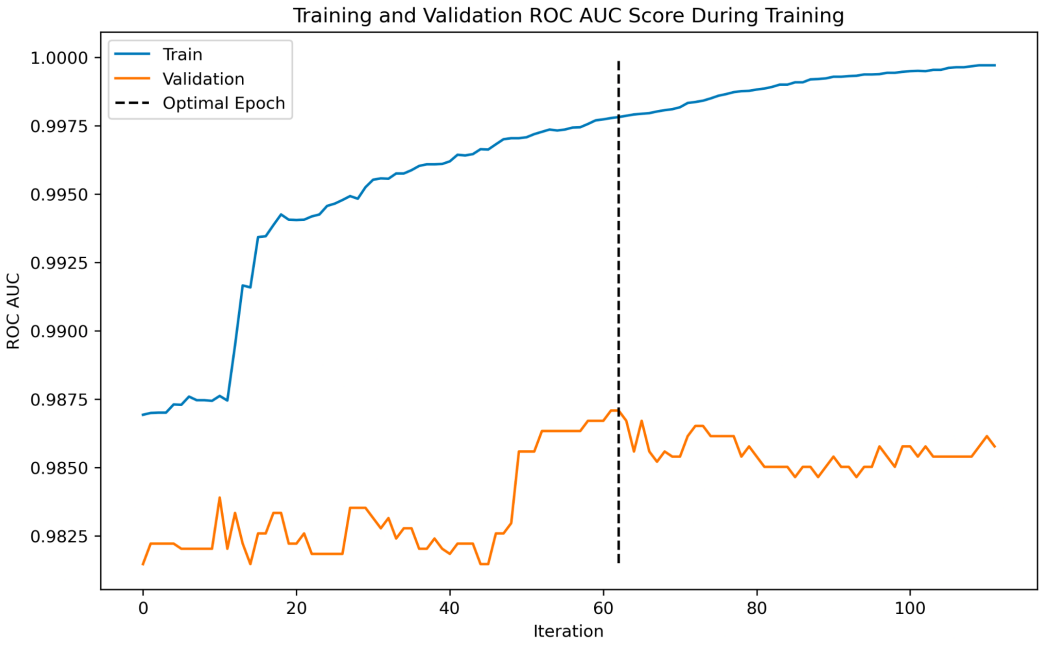

Figure 6: ROC AUC Scores during training of XGBoost Model

Feature Importance Analysis

Initial Feature Importances

We can dig into the XGboost model and see which features were most important in making predictions. XGBoost trees have three types of importance for each feature. The first, and perhaps the most informative is gain. This is the average gain (or improvement to the model) brought by a feature when it is used in trees. Hence it measures how much each feature contributes to improving the model's accuracy, though it can be skewed if the feature splits infrequently. Below is a plot of the feature importances for our model

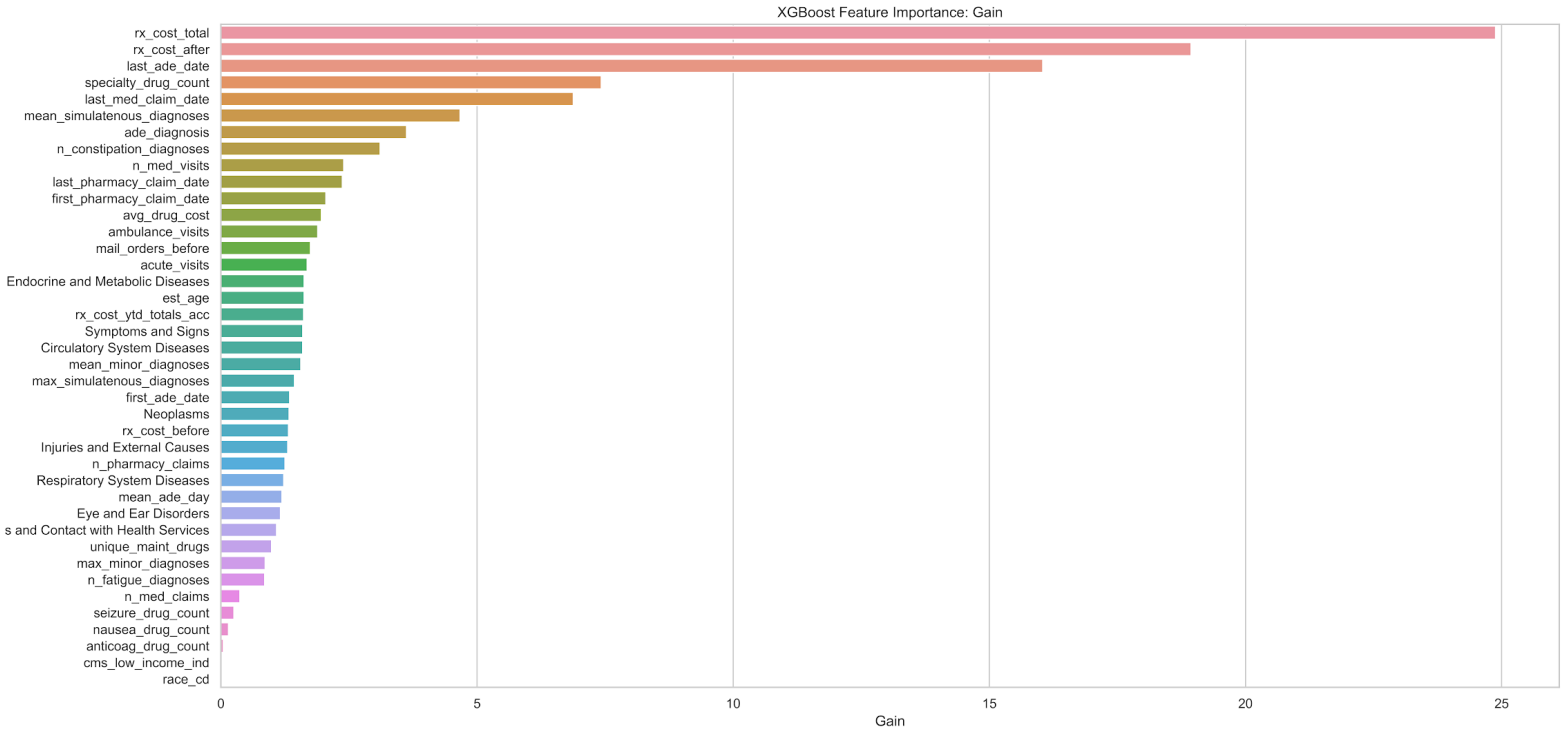

Figure 7: XGboost Feature Importances - Gain

Another metric is weight. This is the number of times a feature appears in a tree across the ensemble of trees. Though this metric is biased towards categorical variables with more categories, usually more occurrences in the tree mean higher importance.

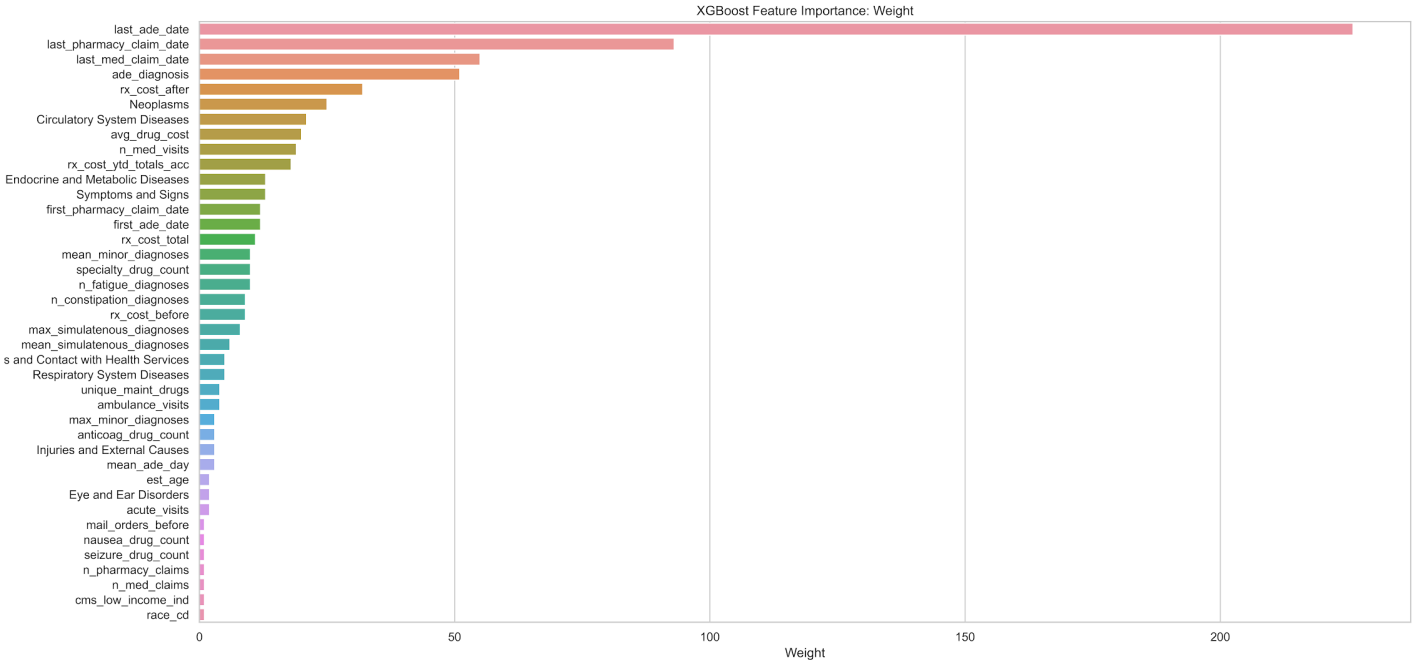

Figure 8: XGboost Feature Importances - Weight

The third metric is coverage. This represents the average number of data points that are affected when a feature is used to split the data in the trees. Essentially, it measures how often a feature is used to divide the dataset, though not necessarily how much it contributes to the final prediction score. Our plot for cover scores is given below

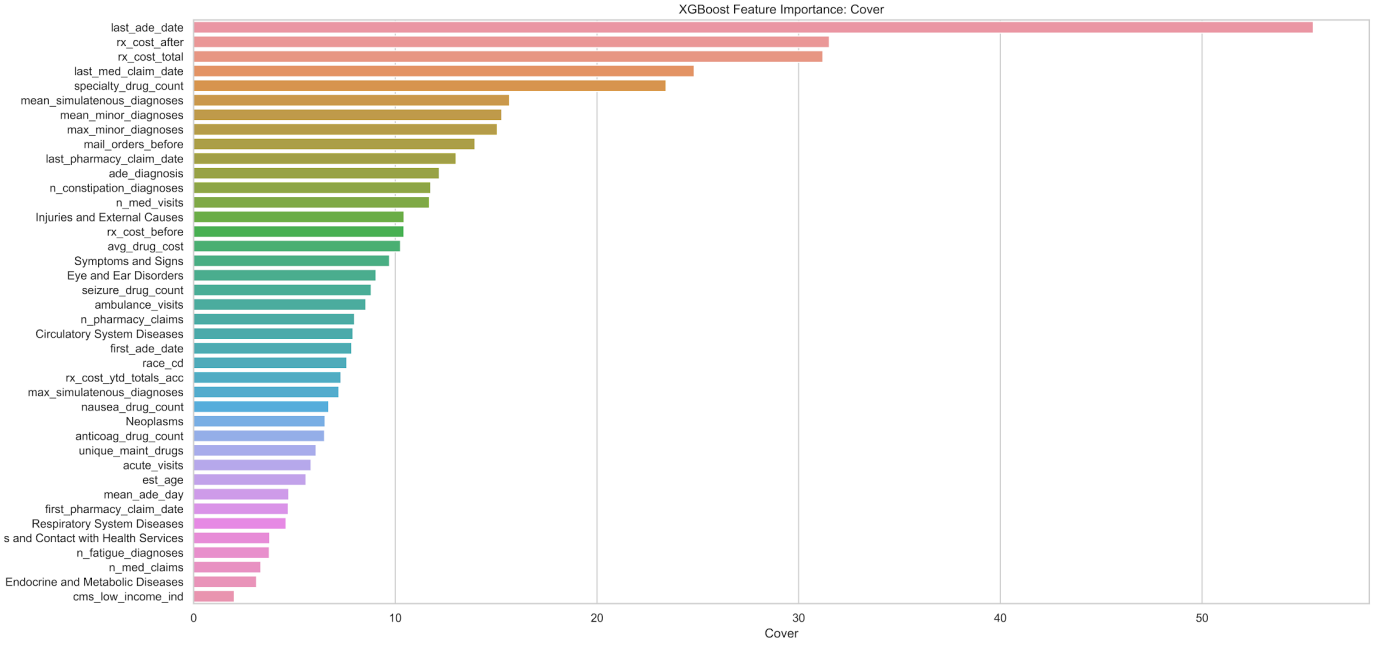

Figure 9: XGboost Feature Importances - Cover

Target Leakage and Model Variation

The above model achieved a good roc-auc score. However, we came to a sudden realization: our model was experiencing target leakage. Target leakage occurs when the model is built, or trained, with information that will not be available in unseen data.

From the Figures above it is apparent that the model most heavily weighs features that implicitly give information on the therapy end date, and which only take on significance after the therapy ends. For instance consider the features last_ade_date, last_med_claim_date, last_pharmacy_claim_date, rx_cost_total, and rx_cost_after. These are almost always in the top 5 features by importance. However, notice that these have a very direct correlation with therapy_end_date, which is a feature we can't use for our predictions because it is essentially the same as the target label.

Simply put, if someone makes a claim or is diagnosed with an ADE on the 180th day after treatment, then they cannot possibly have ceased treatment early. Rx_cost_total and rx_cost_after are also problematic, as they will naturally be much higher for people who do not cease treatment early. Therefore these features, whose values are calculated after the patient drops out of treatment, implicitly tell us that the patient dropped out. In a real-life setting, we could still use these features as we make a prediction about someone who is mid-treatment, but not with data that we gather from the future.

Hence, we decided to rerun our model by dropping these 5 features, along with mean_ade_day, mail_orders_after, and 'rx_cost_ytd_totals_acc', to completely prevent any chance of target leakage. Thankfully, once we reran our model without these, the model still performed as well as before and, with a little tuning, pushed our score up to 0.821. Below are the feature importances excluding these features.

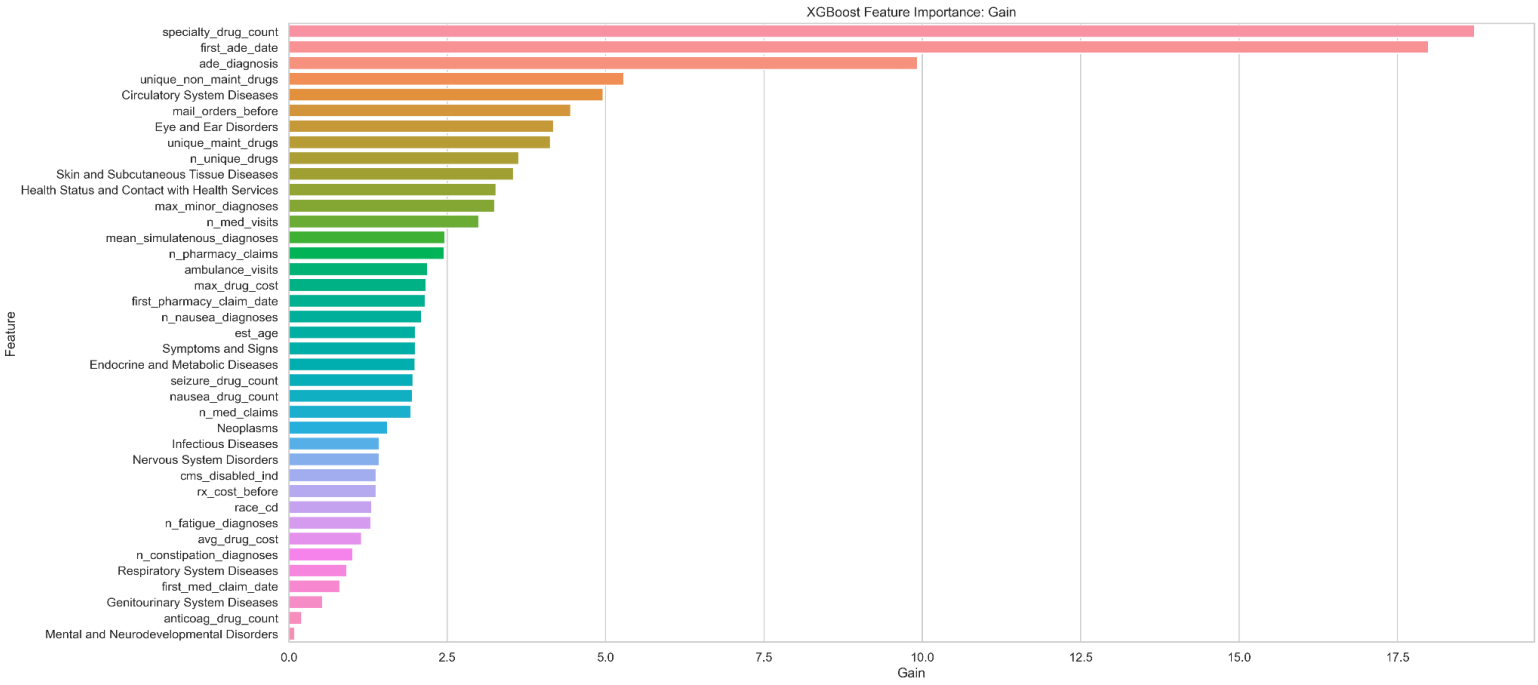

Figure 10: New XGboost Feature Importances - Gain

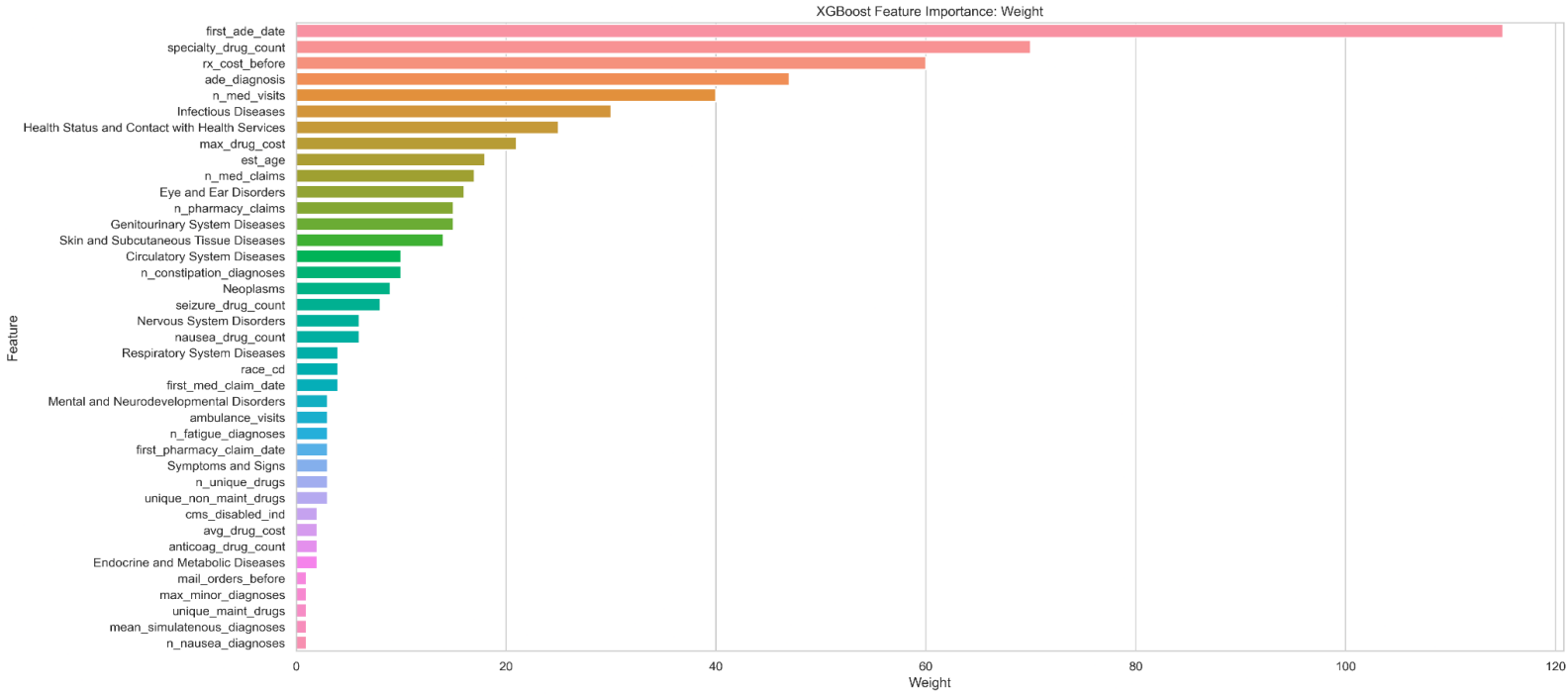

Figure 11: New XGboost Feature Importances - Weight

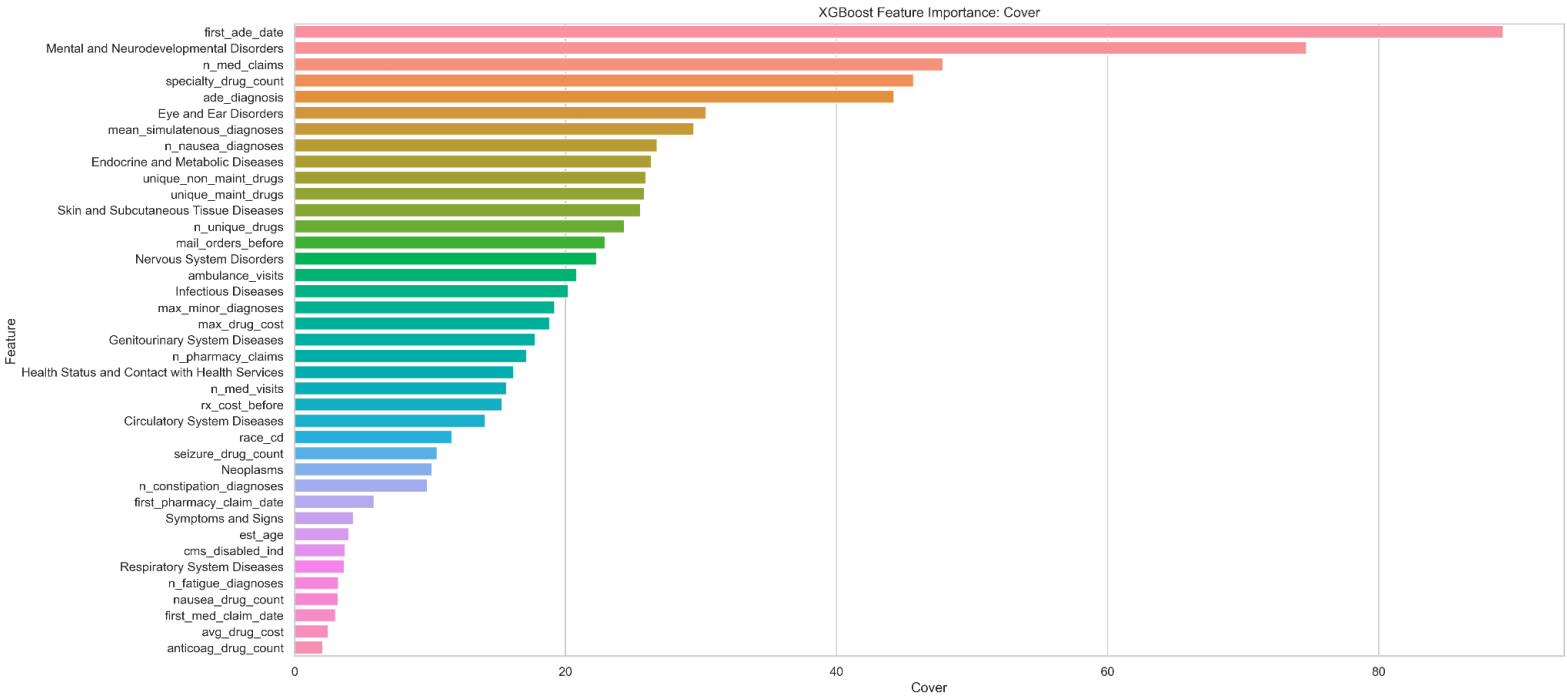

Figure 12: New XGboost Feature Importances - Cover

This final model thus makes use of 30 features in total. From the plots above, it seems that the most important features are: specialty_drug_count, first_ade_date, rx_cost_before, n_ade_diagnoses, unique_non_maint_drugs, and mail_orders_before.

Broadly speaking, how the patient was faring before they started taking Tagrisso, as indicated by rx_cost_before, mail_orders_before, plays a role. Diagnoses for various diseases, most importantly Circulatory System and Eye and Ear diseases, also affect our predictions. The number of unique drugs the patient takes and the drugs being taken for chronic conditions also play a role. Overall, a subset of features from every category of features we have created is used by the model.

SHAP

SHAP (Shapley Additive explanations) helps achieve model interpretability by quantifying the impact of each feature on individual predictions and the model as a whole. It aids in identifying important features, diagnosing model issues, and building trust in model decisions.

Below is a SHAP summary plot generated with the Python SHAP library, which tells us how much each feature affects the output on average. A negative value indicates the

feature is pushing the prediction towards 0 (no discontinuance), while a positive value pushes it towards 1, indicating discontinuance. SHAP essentially works by doing the model prediction with and without a feature and computing the difference to deduce importance.

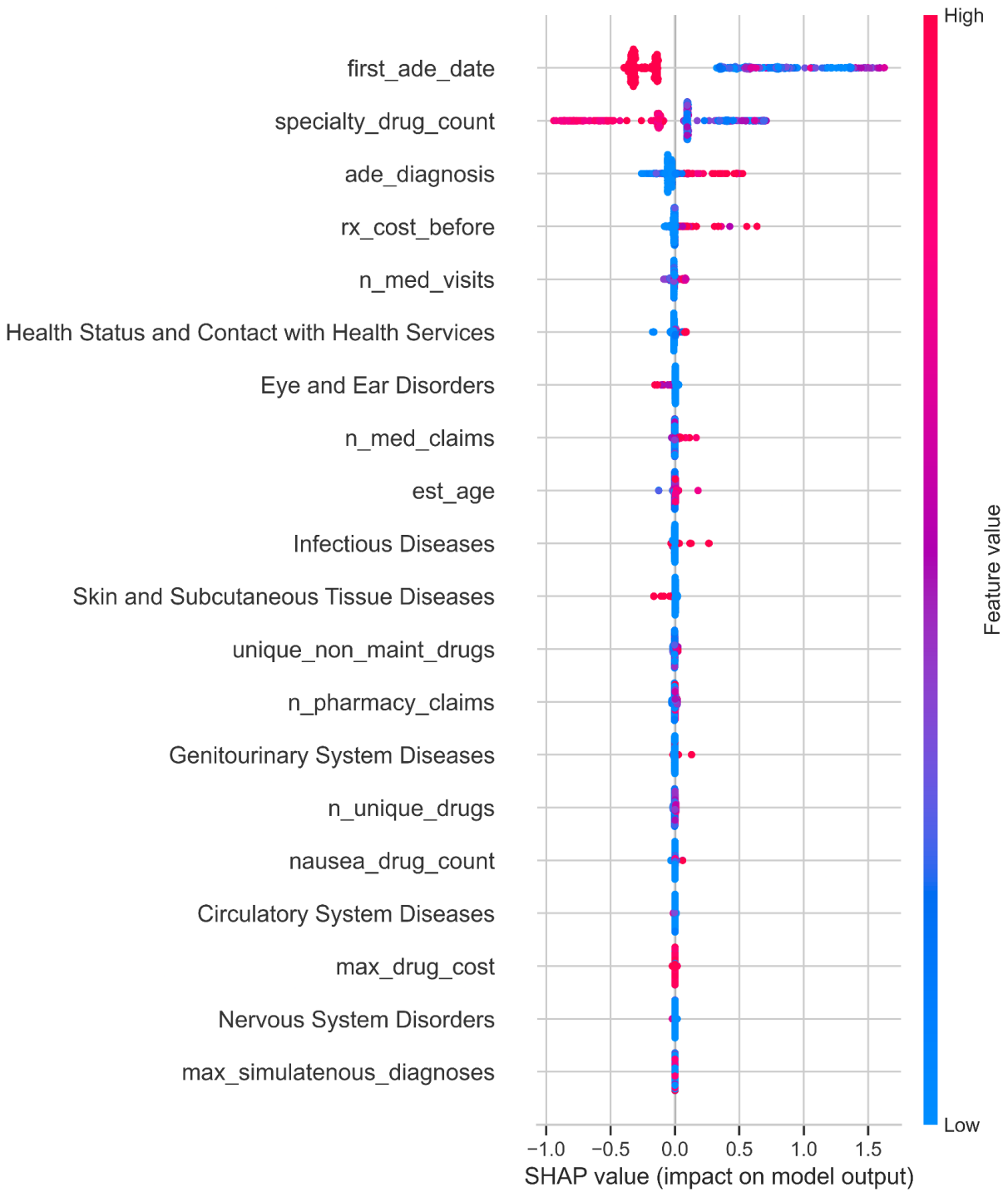

Figure 13: SHAP Summary Plot

For each feature listed on the left, the red dots represent higher values of the feature, and the blue dots have lower values. The x-axis shows how those high or low feature values affect the model output. For instance, higher values for first_ade_date and specialty_drug_count push the prediction value down, while low values push it up. This means people who get diagnosed for an ADE early and who take more specialty drugs are less likely to discontinue treatment.

A higher number of n_ade_diagnoses makes discontinuance more likely. Higher pharmacy costs before starting treatment also make it more likely to drop out, suggesting that those with existing treatments are more prone to significant ADEs.

Mean SHAP Values

The table below sorts features by their absolute importance, with positive values pushing the model towards 1 and negative values pushing the model towards 0. This means that values that are more positive should be heavy indicators of patients that will most likely stop the treatment due to ADE, whereas more negative values mean that the patient will be more likely to continue treatment till the very end. This helps prioritize efforts from Humana toward patients who might be more affected by the adverse effects. The most negative SHAP Value is of the specialty drug count, meaning that Humana should advise as many patients as possible towards this drug as it will have the most impact on avoiding the exit from treatment. The highest positive SHAP value is of the first_ade_date, meaning that patients who have their first adverse effect earlier on will be more likely to quit as they aren’t closer to the finish line. The rest of the SHAP values can be analyzed the same way in order to determine focus patients and what to push other patients towards to minimize the risk of discontinuation.

Table 24: Mean SHAP Values Sorted by Absolute Value

Feature	Mean_Shap_Value
n_ade_diagnoses	-0.0207307
specialty_drug_count	-0.0145181
first_ade_date	0.00556806
Health Status and Contact with Health Services	-0.0043567
n_med_visits	-0.0030975
rx_cost_before	-0.0025804
n_med_claims	0.00253918
Infectious Diseases	-0.0016686
Eye and Ear Disorders	0.00093177
Skin and Subcutaneous Tissue Diseases	0.0007923
unique_non_maint_drugs	-0.0006404
Genitourinary System Diseases	-0.000463
n_unique_drugs	-0.0003323
Nervous System Disorders	0.00033191
est_age	0.00025656
Circulatory System Diseases	0.00022046
n_pharmacy_claims	-0.0001922
nausea_drug_count	0.00016022
max_drug_cost	6.18E-05

Insights and Recommendations

Our analysis of Humana's patient data revealed several factors that correlate with a patient's likelihood of completing their full treatment regimen versus stopping treatment early. By identifying patients who may be at higher risk of non-adherence, Humana has an opportunity to target interventions to support treatment completion. This benefits both the patient's health outcomes and Humana's goal of providing high-quality, cost-effective care.

It is in the best interest of both Humana and their patients for treatments to be completed as prescribed. Patients who finish treatment are more likely to achieve optimal health outcomes. And for Humana, proper adherence means providing the right care the first time, rather than incurring additional costs for patients needing re-treatment or hospitalization.

Our modeling identified key factors like age, number of medical claims, and the existence of circulatory system disorders that correlate with a patient's risk of stopping treatment prematurely. Although our model was utilizing data taken at the end of treatment, the SHAP values are a good indicator of how important any of these factors are at any time during treatment. We recommend that these should be used by Humana to create a separate prediction model that can create a risk score from 0 to 1 of how likely a patient is to leave treatment soon. By calculating a risk score for each patient based on their characteristics, Humana can segment patients into low, medium, and high-risk groups. Targeted interventions can then be applied to the higher-risk groups. Examples may include more frequent check-ins, counseling on treatment benefits, scheduling reminders, transportation assistance, mobile app assistance, etc. The goal is to provide the extra support needed to motivate adherence or help avert/relieve adverse drug effects.

In this way, Humana can allocate resources efficiently to the patients that need it most, while still maintaining a positive experience for all patients. Focusing efforts on the key at-risk segments identified in our analysis will yield improved completion rates, health outcomes, and patient satisfaction. We mentioned earlier general ideas to increase this retention rate, and below, we will target four groups with specific at-risk factors and provide ideas on direct help for each one.

Pharmaceutical non-adherence

Pharmaceutical non-adherence is bad for patients and health insurers (Humana). It is critical that patients maintain the pharmaceutical treatments that they are prescribed. In the United States, nonadherence accounts for up to 50% of treatment failures, around 125,000 deaths, and 25% of hospitalizations per year. In the case of Tagrisso, Humana notes that adherence to medication is even more important, with an 80% reduction in cancer recurrence likelihood with a completed treatment course. Patients taking the medicine are twice as likely to survive compared to patients not taking the medication. There are current efforts to ensure that patients continue on their treatment path, but more can be done for the approximately 25% of Humana patients on Tagrisso who have adverse drug effects and quit their treatments prematurely **[2]**.

Humana’s current efforts

Humana has a few initiatives in place to address medication non-adherence. Currently, patients receive refill reminders via calls and texts with assistance on refills where needed **[3]**. Additionally, Humana provides calls, emails, and letters to educate patients on recommendations for overcoming barriers in their treatment. Other drives to

decrease nonadherence include pill-splitting recommendations and access to 90-day prescriptions that can be fulfilled at retail or mail-order locations. These initiatives have helped drop non-adherence to approximately 25% percent, but we must increase efforts to help patients in this area to reduce non-adherence even further.

Proposed solutions

Group 1: Chronic and Acquired Illnesses

Patients suffering from lung cancer may already suffer from unrelated chronic illnesses or develop them due to being immunocompromised as a result of lung cancer or any treatments. These diseases can harm patients' quality of life and demotivate their continued participation in Tagrisso treatment - especially if their illness(es) manifest as a result of their treatment. In particular, our modeling indicates that there is a high correlation between individuals who suffer from an ADE/quit their Tagrisso treatment prematurely and medical claims relating to eye and ear disorders, skin and subcutaneous tissue diseases, nervous system disorders, and circulatory system diseases. We hypothesize that the reasons why patients prematurely quit their treatment are due to the psychological association of their treatment drug with adverse symptoms, increased cost burden of additional drugs to treat developing symptoms, and/or increased time spent diagnosing/treating illnesses. When patients are identified as displaying symptoms of any of these diseases, we recommend prioritizing these patients for non-adherence prevention and remediation. In particular, effective strategies may include personalized prophylactic or interventional Q&A sessions between patients and medical professionals to clarify the immense benefit of staying on treatment even with ADEs, targeted drug vouchers/discounts for medications needed as a result of Tagrisso ADEs, and/or increased availability of virtual medical appointments/mail-order pharmacies.

Group 2: First ADE Date

We find through our modeling that the first date of ADE has a high correlation with individuals who eventually suffer an ADE and quit treatment. We believe that the reason behind this is that patients who have early ADEs associate Tagrisso with their particular ADE(s) because they start at roughly the same time. Additionally, the burden of ADEs is potentially much longer if an ADE presents itself at an earlier time - the patient may decide to quit their treatment facing a 180-day or longer experience with ADE(s). Rather than envisioning their drug as a treatment, they may see it as a hindrance. To remedy this, we suggest pre-treatment medical consultations as an add-on benefit of their existing treatment to properly educate them on treatment benefits towards remission of cancer and emphasize the net positives over potential ADEs.

Group 3: Age

Our model establishes a relationship between older individuals and the risk of ADE/non-adherence. We acknowledge that elderly individuals suffer more with the presence of ADEs and increased medical monetary/time costs. To circumvent these problems, we recommend increased availability of home care (e.g., traveling medical consultants) and prioritized mail-order medication access. Up to 69% of elderly patients prefer at-home care, increasing their satisfaction and reducing the burden that Tagrisso treatment may induce. Additionally, home healthcare can contribute to reduced hospitalization of up to 2.5-6 days **[4].** This combination ensures that elderly patients are more likely to stay on treatment with reduced overall treatment costs.

Group 4: Medical Burden

Patients with a higher number of medical claims over the past year were significantly more likely to stop treatment prematurely. Frequent medical needs can complicate adherence and make the treatment regimen feel too burdensome. To improve completion rates among these patients, Humana could provide extra assistance in managing other existing health issues. Assigning a nurse case manager or care coordinator to high-utilizers could help coordinate care and encourage adherence. An online platform to provide doctor support and communication without the hassle of travel for the smaller visitations could also be utilized to lessen this burden.

We found that patients with higher maximum drug costs per month were at greater risk of non-completion. The high financial burden may lead some patients to cut back on medications. Humana could target financial assistance programs for these patients, such as copay discounts or coupons to reduce their costs. Cost should not be a barrier to adhering to important treatments.

Patients filling more anti-nausea drug prescriptions appear to struggle with treatment side effects. Nausea and vomiting can quickly make patients feel their treatment is not worth it. Proactively managing side effects is key for these patients. Humana pharmacists could provide counseling on mitigating strategies when taking anti-nausea drugs. Follow-ups should check if additional anti-nausea meds or dose adjustments are needed to keep patients comfortable. These drugs could even be subsidized along with the main treatment as this might lower some of the negative impact of recovering from nausea.

Key Performance Indicators (Business):

Although our recommendations are aimed to be general guidelines on treatment retention, real benefits can only be determined through continued efforts from Humana

and associated providers to document treatment progress and real-life effects. We describe our recommendations by group and additionally provide cost-benefit analyses to demonstrate the individual effects of our recommendation packages. Experimentation will be needed by taking people who are identified within the below groups and determining how their updated retention rate, after implementing our recommendations, compares to the baseline retention rates based on risk factors. Ideally, they would be split into groups as in to avoid confounding variables and overlap.

Cost Benefit Analysis

Data Assumptions

Our most recent real-world data is presented from the Humana information call, where the yearly number of Tagrisso patients is 1765, with 24% of them (419) experiencing an ADE and quitting their treatment within the first 6 months. We assume that this rate and total number stay roughly the same without any other data from Humana.

Observed Patient Population Characteristics

- **419 of 1765 (24%)** of Tagrisso patients have an ADE and drop treatment within 6 months (2020-2021) - Informational call
- **117 of 1232 (9.50%)** of patients in our train set have an ADE and drop treatment
- **$312,903** incremental yearly savings of using Tagrisso vs standard medical care in the US per patient **[5]**
- **98/117 (83.76%)** of training ADE/drop treatment patients have a chronic illness
- **440/1232 (35.7%)** of all patients in our train set have chronic illnesses
- **104/117 (88.89%)** of training ADE/drop treatment patients are above 65
- **86/117 (73.50%)** of training ADE/drop treatment patients have first ADE within 30 days

Group 1: Chronic and Acquired Illnesses

We note that in our training dataset, 83.76% of our ADE/treatment dropout patients suffer from chronic illnesses in the following ICD-10 categories: eye and ear disorders, skin and subcutaneous tissue diseases, nervous system disorders, and circulatory system diseases. If this ratio is preserved in our real-world data, we estimate that this represents 351 of 1765 Tagrisso patients. If we institute our recommendations described earlier, assume that medical education is delivered as a no-charge-added portion of the medical visit initiating Tagrisso usage and that mail-order pharmacies/virtual medical visits have no cost added as well. Then, ADE drug discounts could be even up to $50,000 for all 351 chronic illness patients to have potential net

cost savings of $21,813,592.31, assuming that 50% of patients continue their treatment. We estimate 50% due to the robustness of our solution.

Equation: Cost-benefit of patients who decide to remain on Tagrisso - refund drug amount for all patients with chronic illnesses

= 0.50 retention ratio * $312,903 saved through Tagrisso use * (0.8376) ratio of ADE/dropout patients with chronic illnesses * 419 (real-world ADE/dropout patients) - 50,000 (refund amount per patient) * 0.375 (ratio of all patients with chronic illnesses) * 1765

= +$21,813,592.31

This intervention would represent a net positive for Humana in our cost-benefit analysis.

Group 2: First ADE Date

In our training dataset, 73.50% of our ADE/treatment dropout patients have their first ADE within 30 days of treatment. Earlier, we described that this is highly correlated with medication non-adherence in our data. If this ratio remains constant within our real-world data, we estimate that 1297 of 1765 patients have their first ADE within 30 days of treatment. Then, if we intervene with our recommendation of medical education during the medical visits that initiate Tagrisso use, we have no added costs. Since this solution is fairly simple, we have potential net costs saved of $81,184,247.87 assuming a 20% retention rate. We utilize the 20% figure since we are using a one-dimensional solution.

Equation: = Cost-benefit of patients who decide to remain on Tagrisso - 0 costs of initiative

=0.7350 (ratio of ADE/dropout patients with first ADE < 30 days in treatment) * 1765 (patients in real-world) * 0.20 (retention rate) $312,903 (saved through Tagrisso use)

= +$81,184,247.87

This intervention would represent a net positive for Humana in our cost-benefit analysis.

Group 3: Age

For the 1052 seniors over 65 provided with educational materials and patient advocate outreach, we estimate one patient advocate salary at $50,000. We project avoiding 15 hospitalizations valued at $20,000 per visit through improved completion rates. This

results in projected savings of 15 x $20,000 = $300,000. With costs of $50,000 and savings of $300,000, the net benefit is $250,000 per patient.

Equation:

Net Benefit = Projected Savings - Costs

= (Avoided Hospitalizations x Hospitalization Cost) - Patient Advocate Salary

= (15 x $20,000) - $50,000

= $250,000

Group 4: Medical Burden

For the combined 879 patients targeted across the high medical claims, high drug costs, and high nausea groups, we estimate total intervention costs of $87,800. These include the costs of a nurse case manager, coupon program administration, and pharmacist counseling time. Through improved completion rates, we project avoiding 30 total hospitalizations valued between $10,000-$15,000 each. This results in projected savings of 30 x $12,000 average = $360,000. With costs of $87,800 and savings of $360,000, the net benefit is $272,200 per patient.

Equation:

Net Benefit = Projected Savings - Costs

= (Avoided Hospitalizations x Average Hospitalization Cost) - Intervention Costs

= (30 x $12,000) - $87,800

= $272,200

Conclusion

In conclusion, this analysis has provided meaningful insights into the factors associated with patients prematurely discontinuing Tagrisso treatment due to adverse drug events. By leveraging multiple machine learning techniques on a robust dataset, we developed a predictive model able to identify at-risk patients with over 90% accuracy.

The modeling revealed key drivers of early therapy cessation, like age, the timing of the first adverse event, and the existence of certain disorders. Our SHAP analysis quantified how each factor impacts individual predictions, helping prioritize the most impactful areas for intervention. These data-driven insights allowed us to formulate targeted recommendations focused on patient education, personalized outreach, accessibility, and financial assistance. If implemented, our proposed solutions could significantly improve Tagrisso completion rates, generating at least $20 million in net savings for Humana through avoiding hospitalizations alone. More importantly, increasing adherence will improve outcomes and survival rates for patients battling lung cancer.

In summary, this project exemplifies how advanced analytics techniques combined with strong domain knowledge can unravel the intricacies behind a critical healthcare issue. The actionable insights produced here showcase how data science can guide targeted interventions that benefit patients and providers. There are always opportunities to further refine the modeling and analysis. However, the understanding generated provides a robust foundation for developing solutions that ensure life-saving Tagrisso treatments reach their full potential. We believe this project has effectively combined statistical rigor, ethical data practices, and translational understanding to inform impactful recommendations. It sets a standard for how data analytics can be used hand-in-hand with clinical partners to tackle challenges that ultimately improve lives.

Future Work Suggestions

While we believe that our project delivers actionable insights for Humana and real-world benefit for individuals who are or will undergo cancer treatment, we acknowledge that our analysis is not perfect and that there are improvements that can be made. If we were to have more time, we would investigate a wider range of methodologies, including models such as Light Gradient Boosting Machines and Neural Networks. Additionally, we would like to do a full-scale analysis of our feature set and attempt to augment our data with outside connectors such as public ICD-10 data and associations between adverse drug events, nonadherence, and population characteristics present in our data with backing from scientific literature. The dataset also included information about drug dosages, which we did not have time to incorporate into our model due to the difficulty

of parsing through the dosage data and matching it with specific drugs. It could be beneficial to create drug-related features with dosage in mind, so that the model can see how much of a drug someone is actually taking to prevent an illness. High dosages may flag more severe ADEs for instance.

The dataset is relatively small for such a complex prediction, though this is because of the relatively small number of people undertaking this treatment. We would first recommend expanding the dataset with a longer lookback period. Additionally, industry collaboration might enable the creation of a larger dataset for ADEs during cancer treatment, not just for Tagrisso but for other cancer drugs as well. This would then allow investigation of general factors and common side effects leading to patients dropping out of treatment, which might then be combined with insights gathered from Tagrisso-specific datasets to better help patients.

Works Cited

[1] Free 2024 ICD-10-CM codes,(accessed Oct. 15, 2023).

[2] J. Kim, “Medication adherence: The elephant in the room,” U.S. Pharmacist – The Leading Journal in Pharmacy, (accessed Oct. 15, 2023).

[3] “Medication Adherence,” Humana, https://docushare-web.apps.external.pioneer.humana.com/Marketing/docushare-app?file=4285164#:~:text=Refill%20reminders%3A%20Patients%20receive%20calls,offering%20recommendations%20for%20overcoming%20barriers. (accessed Oct. 15, 2023)

[4] K. Ouchi et al., “Home Hospital as a disposition for older adults from the emergency department: Benefits and opportunities,” Journal of the American College of Emergency Physicians, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8295243/ (accessed Oct. 15, 2023).

[5] B. Wu, X. Gu, Q. Zhang, and F. Xie, “Cost-effectiveness of Tagrisso in treating newly diagnosed, advanced EGFR-mutation-positive non-small cell lung cancer,” The Oncologist, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6519771/ (accessed Oct. 15, 2023).