2023

Humana-Mays Healthcare Analytics 2023 Case Competition

PREDICTING FACTORS LEADING TO DISCONTINUATION OF NSCLC TAGRISSO THERAPY

Table of Contents

Executive Summary:

Each year, cancer is responsible for around 600,000 deaths in the US. One of the most common types of cancer is Lung Cancer and it affects more than 200,000 people each year in the US. It is responsible for around 1 in 5 out of all cancer related deaths in the US, making it the leading cause of cancer deaths in the country. In recent years, highly effective and advanced targeted therapies for treating some specific types of cancer have been introduced. One such therapy covered by Humana is Osimertinib (Tagrisso) - used to treat early-stage EGFR+ non-small cell lung cancer (NSCLC). This therapy increases the likelihood of a patient's survival by two-fold and reduces the likelihood of recurrence by 80%. However, these therapies are usually associated with several side effects that are usually manageable. Humana aims to help its members that are undergoing this treatment in successfully continuing and completing the treatment. Our overall objective is to provide Humana with a framework that can be leveraged to proactively identify patients at risk of leaving and suggest solutions to help these patients successfully continue the treatment journey.

Throughout our analysis, we conduct a meticulous study of pharmacy and medical claims data to predict and understand occurrences of Adverse Drug Events (ADEs) and the subsequent discontinuation of Tagrisso therapy within the first 180 days of therapy start. However, our goal is also to truly understand a patient's treatment journey, pinpoint the optimal intervention timelines and ensure personalized, equitable care. We begin with an in-depth background research to understand a patient's journey through cancer. This is followed by detailed exploration of data and data cleaning to gain a preliminary understanding of the data and prepare the data for the modeling phase. Three models - Random Forest, Light Gradient Boosting Machine (LGBM), and CatBoost were rigorously tested, along with an exploration of a Meta-Ensemble Stacking Approach. Despite the Stacked Ensemble model demonstrating superior efficacy, the LGBM Model was chosen as the final model (AUC 0.9659, 10-fold CV), thoughtfully balancing predictive accuracy, model simplicity, and interpretability, while ensuring computational efficiency and clear insight into model-driven decisions.

The predictive modeling phase is followed by a study of the predictive importance and directionality of the features. These insights are then utilized to build a segmentation model for grouping the patients into 4 different groups based on their demographic profiles, side effect prevalence, Tagrisso usage, and pharmacy claims history. Post the segmentation, the groups are studied and labelled based on their level of vulnerability which is further leveraged for prioritization of intervention strategies. Our analysis has also been corroborated by an analysis of external data on drug reviews and patient comments gathered from the web. Sentiment Analysis and Topic Extraction on this text data revealed a few key sentiments and associated pain points for the patients that are undergoing the treatment.

Based on the patient segmentation and understanding of the key areas of disconnect, we propose a framework to re-imagine the "Illness to Wellness" patient journey that helps understand the patient's needs at every step of their journey and take proactive actions. We recommend Humana to tackle the problem by adopting a three-pronged strategy –

- Financial support for the less economically privileged groups by assisting with deductibles and coinsurance.
- Ease of access to medical re-consultation and medical support services for at-risk groups through prioritized channels like Home Care and Virtual Care
- A re-imagined and continuously accessible patient support network through dedicated case managers and specialized support groups.

A detailed cost-benefit analysis for our proposed Re-Imagined Illness to Wellness journey, indicates potential revenue savings in the range of $61.7 million, underlining the financial and therapeutic prudence of maintaining patients on Tagrisso therapy for 180 days versus other alternative treatment options. This recommendation acts as a pillar of support for the patient throughout their cancer treatment journey and marries economic sensibility with a deeply empathetic, patient-focused care model.

1. Introduction:

Cancer is the second most common cause of death, exceeded only by heart disease in the United States anticipating approximately 609,820 fatalities in the United States alone in 2023 [1]. This ailment presents not only a physical challenge but also a significant emotional and psychological burden for patients and their families. Compounding this, lung cancer emerges as a particularly lethal adversary, with a staggering 2.2 million people diagnosed globally each year with 80-85% of these diagnoses attributed to non-small cell lung cancer (NSCLC) [2]. This substantial figure not only underscores the pervasive nature of NSCLC but also highlights a critical area of focus for medical research, patient care strategies, and health policy planning.

This report focuses on a specific subtype of non-small cell lung cancer (NSCLC), characterized by the EGFR (epidermal growth factor receptor) mutation. EGFR plays a crucial role in abnormal cellular growth, leading to the progression of the disease. One of the available therapeutic strategies is Osimertinib, an oral tyrosine kinase inhibitor, predominantly utilized in the treatment of early-stage NSCLC due to its demonstrated efficacy. Patients treated with Osimertinib are observed to have an 80% lower risk of cancer recurrence or mortality. However, the administration of Osimertinib is often accompanied by a suite of side effects, including nausea, fatigue, pain, elevated blood glucose levels, and constipation. These side effects, although manageable with proper guidance and auxiliary medication, often lead patients to discontinue treatment prematurely. Osimertinib is branded and distributed as '*Tagrisso*' by AstraZeneca Pharmaceuticals.

Humana, as a premier health insurance provider, commits to supporting individuals through their treatment journey. Beyond just providing financial assistance, this commitment helps patients overcome the difficulties presented by cancer treatment, covering anything from managing side effects to coping with any kind of delays or confusion.

The core objective of our study is to understand the reasons behind patients' discontinuing therapy, especially regarding the manageability of Tagrisso's side effects. Through a nuanced exploration of patient experiences and data, we aim to derive insights that could pave the way for improved patient support and management strategies, ensuring that individuals do not navigate their cancer journey in isolation.

2. Problem Statement:

This analysis focuses on resolving the business problem of identifying members experiencing side effects or Adverse Drug Events (ADEs), leading to therapy discontinuation within 180 days of using Tagrisso. The report does not restrict its scope to the side effects alone but also looks to leverage other likely contributors in their Pharmacy (rxclaims) and Medical claims (medclaims) data. The

predictive model built herein forms the foundation of our proposal, providing recommendations that could enhance Humana's ability to understand and address patient vulnerabilities, enabling them to intervene with appropriate care.

The depth of the problem extends beyond merely identifying why members might withdraw from therapy prematurely; it also explores when they are most vulnerable, empowering Humana to make timely interventions. We understand that patients experiencing ADEs may not be a homogenous group, and therefore our report offers personalized recommendations to appropriate patient segments. Ultimately, our objective is to mitigate any potential bias in the data, ensuring that our analyses and recommendations uphold the principles of fairness and equity.

3. Preliminary Research:

We feel it is imperative to understand the customers' journey in the clinical environment to identify the potential breaking points that could lead to premature therapy discontinuation. We believe that it is extremely important to look at the problem from the patients' lens to suggest relevant recommendations. The emotional and physical burden of cancer is already very daunting and if there are process-related issues on top of that, it could lead to a very frustrating experience for the patient.

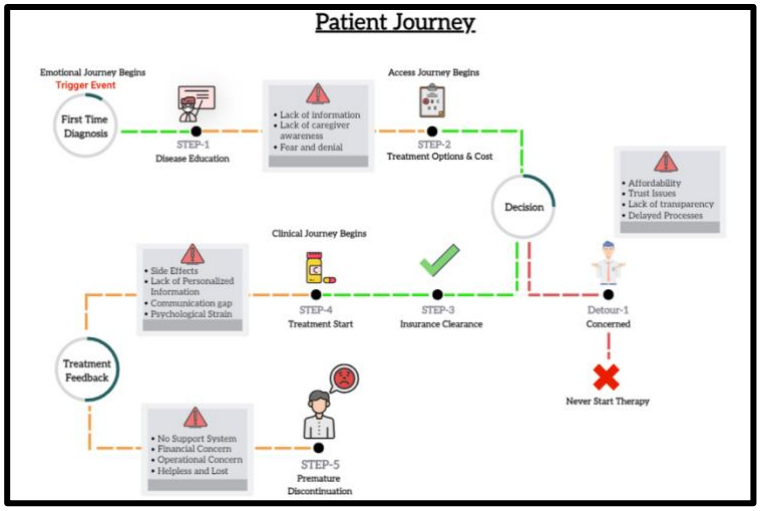


Figure 1: Patient Journey

Figure 1 illustrates the sequential stages of a patient starting from the *'Trigger Event'* i.e., the first-time diagnosis of cancer. The steps and vulnerabilities can be understood below:

- *First Time Diagnosis (Trigger Event):* This is a primary step at which the patient learns that they have been diagnosed with NSCLC.
- *Disease Education (STEP-1):* This step of the journey is where patients are educated about the diagnosis itself and also made aware of how the disease affects individuals. This is a foundational step for the patient and the insurance providers to navigate through the rest of the journey properly. However, a research study approved by Mayo Clinic Arizona in June 2018 found that only 50% of the 113 patients diagnosed with cancer were equipped with knowledge about financial management while a mere 35% with pain management [3]. Additionally, lack of education also builds fear and denial in the patients' minds which if not addressed properly could lead to them behaving purely out

of emotion. Lastly, the lack of awareness among caregivers is also instrumental for the success of cancer therapy as patients sometimes may not be physically and/or mentally able to care for themselves.

- *Treatment Options & Cost (STEP-2):* The plan of action is the next step wherein customers weigh out the pros and cons of considering different therapy options given their financial and physical competence.
- *Decision*: This auxiliary step signifies the patient's decision to move ahead with a chosen plan. But sometimes, there are concerns regarding the way forward. A few examples of these are given below.
    - Affordability: Patients' are concerned about whether they will be able to afford medical care given their current insurance plan. Patients with economic constraints might find it hard to cover out-of-pocket expenses.
    - Trust Issues: Some patients might distrust the medical system or their healthcare provider due to information they read online or receive through word of mouth. It is crucial for the insurance care provider to establish a robust line of communication with the patient else a lack of reassurance could lead them to lose the patient.
    - Lack of Transparency: Patients may have encountered hidden costs in the past, which might make them hesitant to commit to treatment for a difficult condition like cancer.
    - Delayed Processes: During cancer diagnosis, time is of the essence but we often see that patients take a long time to decide the course of treatment. Even when patients make quick decisions, they fall prey to being stuck in the process of paperwork and insurance approvals which could exacerbate their pessimism and lead them to never start therapy.
- *Insurance Clearance (STEP-3):* Once the patients' insurance is approved, they can move on to the next step which is to start the treatment.
- *Treatment Start (STEP-4):* This step is where the patient's clinical journey truly begins in the process of cancer therapy.
- *Treatment Feedback:* The patient is bound to come back to their clinical advisors (medical and insurance providers) to highlight challenges faced during their initial experience of starting the treatment. The patient may face the following challenges leading to discontinuation of therapy:
    - Side Effects: At this step, patients may start showing signs of side effects like Nausea, Fatigue, Constipation etc. which might be disrupting their quality of life. This discomfort can be unbearable in certain sections of the patient population, which could lead to discontinuation of therapy if they are not dealt with properly.
    - Lack of Personalized Information: Adoption of generalized insurance plans may lead the customer to drop out of therapy as they may not be able to afford the treatment due to rising costs from side effect medications, emergency room (ER) visits, increase in expenditure due to fatigue etc.

- Communication Gap: If the service providers are unavailable for the patient during this period of treatment, it elevates their anguish as they are not able to receive proper care for their problems. Patients require someone to hear them out in these circumstances which will help them to move forward with therapy.
- Psychological Strain: Smith et al. (2015)[4] observed that Depression is a common comorbidity in cancer cases, affecting >10% of patients. A meta-analysis also revealed that minor or major depression increases mortality rates by up to 39% and that patients displaying even a few depressive symptoms may be at a 25% increased risk of mortality.

- *Treatment Discontinuation (STEP-5):* Finally, even after the treatment feedback, patients may still face financial and operational challenges concerning their medical and insurance providers, leading to therapy discontinuation. The feeling of loneliness due to the lack of a support system makes them feel helpless during the cancer treatment. If they are not heard and their issues are not addressed at the right time, they most certainly will feel the need to explore other options.

4. Understanding the Data:

A. Data Overview:

We have been given three different datasets (both for training and holdout) to analyse, predict, and understand the patients on the Osimertinib (Tagrisso – 1) medication who are leaving the treatment within 180 days (or roughly 6 months) from the start of the treatment and have reported an ADE. The following is a brief description of the three datasets that we received:

- Target dataset (contains 1232 patient records in the training data sample and 420 patient records in the holdout data sample): The data is present at the unique individual patient and therapy level with the information of the therapy start date. This dataset contains demographic information for the patient such as age, sex, race etc. that enabled us to make sure that our predictive model is unbiased and does not favor any privileged group. This data set also contains the target variable ("tgt_ade_dc_ind") that if equal to 1 signifies that the patient has left the treatment and reported an ADE within 6 months of the start of treatment.
- Medical Claims dataset (contains 100159 claims in the training data sample and 23232 claims in the holdout data sample): The data is present at a unique medical claim level with information on the place of treatment for the patients along with primary and non-primary diagnosis codes for each of the individual medical claims. This dataset also contains information about an ADE concerning Osimertinib for each of the individual medical claims.
- Pharmacy Claims dataset (contains 32133 claims in the training data sample and 6670 claims in the holdout data sample): The granularity of the data set is again at the pharmacy claims level with the information of different drug classes and the dosage of the drugs for each claim. This

data also contains indicators of specific drugs that are known to cause an ADE in someone who is on Osimertinib medication.

The timeline of the data (for both medical claims and Pharmacy claims) present for an individual starts from a maximum of 90 days before the patient begins the Osimertinib therapy through the end of their therapy.

Initial Data Exploration:

To understand the basic patterns in the data, we wanted to explore the different datasets and then use that information to process the data so that it is fit for predictive modelling.

1. *Distribution of the target variable (tgt_ade_dc_ind):*

   We notice that the distribution of the target variable is slightly imbalanced. The positive class (tgt_ade_dc_ind = 1) is at 9.5% compared to the negative class (tgt_ade_dc_ind = 0) at 90.5%. The positive class of the target variable signifies those patients have left the treatment and reported an ADE. Most machine learning models prefer a uniformly distributed target variable for their optimum performance and therefore we are going to apply appropriate steps in the model-building methodology to adjust for this.

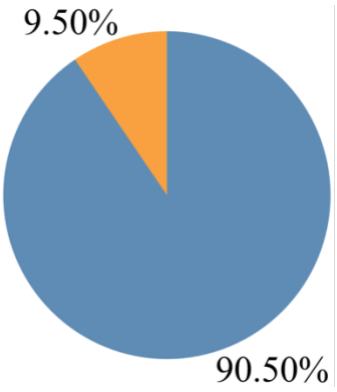


Figure 2: Target Variable Distribution

2. *Distribution of the target variable for the protected variables:*

   For this entire preliminary analysis we have not treated the data in any way, nor we have imputed the null values.

   We notice that the distribution of the target variable with age is mostly similar apart from the lower tail of the box-whisker plot for the positive class. The median age for both the positive and the negative classes differs by 1.

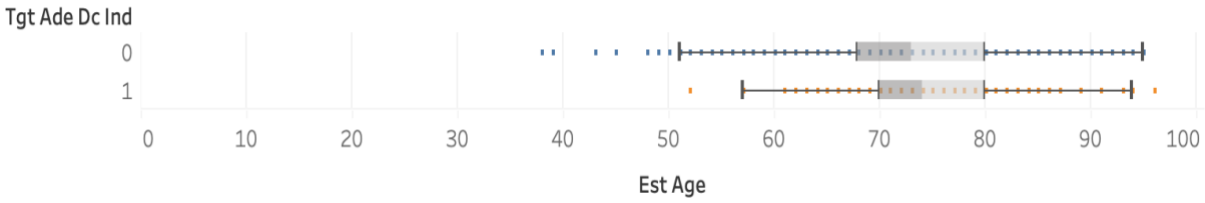


Figure 3: Boxplot of Age with Target Variable

Next, we analyze the "cms_disabled_ind" flag present in the target data. We see that across the positive and negative classes in the target variable, the "cms_disabled_ind" flag has overwhelmingly 0 value. This suggests that most of the patients that we are analyzing are not classified as disabled by the Centers for Medicare & Medicaid Services (CMS).

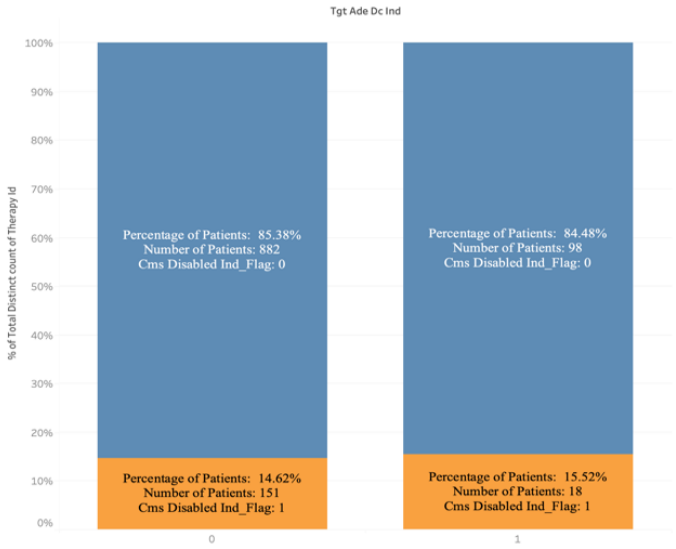


Figure 4: Disability Indicator with Target Variable

After this, we analysed the patients who are receiving low-income subsidies from CMS and what the proportion looks like with the target variable. We can see that 37.07% of patients who are leaving the treatment are on low-income subsidies. On the other hand, 39.01% of patients who are not leaving the treatment are on low-income subsidies.

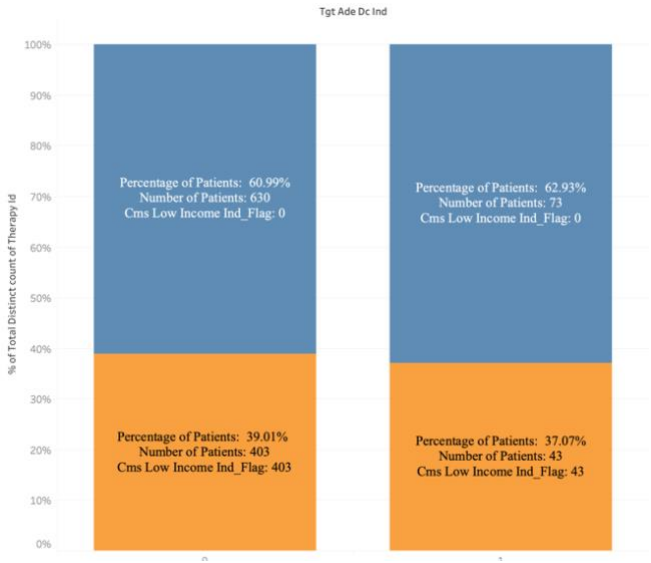


Figure 5: Low Income Indicator with the Target Variable

We looked at the distribution of the target variable with the different Races in the "Race_Cd" variable. We noticed that the highest percentage of people belong to "Race_Cd" = 1 race amongst the patients who are leaving and continuing the treatment. We realized that patients from "Race_Cd" = 6 race is the least across the two categories of the target variable.

Figure 6: Race with the Target Variable

We see that on average each patient has filed around 25 pharmacy claims and around 171 medical claims.

The following table showcases the percentage of people across the training and holdout data set who have filed Pharmacy claims for the different drugs per the specific indicators given.

What is the claim for?	Percentage of People
An Anticoagulant	5.16%
A drug used to treat diarrhea	0.66%
A drug used to treat nausea	3.27%
A drug used to treat seizures	0.67%

Table 1: Percentage of Different Pharmacy Claims

The table below showcases the percentage of people across the training and holdout data set who have filed Medical claims for the different diagnoses per the specific indicators given.

What is the claim for?	Percentage of People
Seizures	0.29%
Pain	0.71%
Fatigue	2.89%
Nausea	2.17%
Hyperglycemia	0.22%
Constipation	1.39%
Diarrhea	0.94%

Table 2: Percentage of Different Medical Claims for ADE

While doing the initial data exploration, we noticed that there are 72 patients (in the Target training data set) whose information is not present in the training data set for Pharmacy claims. For medical claims, we noticed that there are 696 patients whose information is present in the Target training data set but not in the Medical claims training data sets. When we combined the two data sets, we noticed that there were 59 patients whose information was not provided in either of the Pharmacy claims or Medical claims data sets. For the holdout data, we saw a similar pattern. We noticed that 41 patients had no data for Pharmacy claims and 235 patients had no records for Medical claims when compared to the Target holdout data. Again, when we combined the Pharmacy and Medical claims holdout data, we noticed that there are 31 patients whose any kind of record does not exist in these two data sets but are mentioned in the target holdout data.

B. Data Preprocessing:

Data Imputation:

We realized that across the training and the holdout data, the patients whose information is not present in either Pharmacy claims or Medical claims belong to the negative target class (tgt_ade_dc_ind = 0). This tells us that the people who are not filing any kind of claims are probably not facing any ADE (or they can treat the ADE by themselves) and therefore not leaving the Tagrisso

treatment. Now, to appropriately identify the patients who do not have the information in the Pharmacy claims data, but whose information exists in the Medical claims data we put zeroes for the ADE indicator variables under the aforementioned assumption. A similar logic was followed for the patients whose data was available in the Medical claims but not in the Pharmacy claims. Along with this, we filled the columns with the information about primary diagnosis ("primary_diag_cd") with zero as well. Now, for the patients who do not have any Pharmacy claim or Medical claim records, we filled the columns for those two data sets with 0. For the protected attributes we introduced indicator variables and imputed the Age ("est_age") with the average age of 73.771976, and Race ("race_cd"), Sex ("sex_cd"), Disabled ("cms_disabled_ind") and Low Income ("cms_low_income_ind") with -1 to signify that the data was missing.

Data Aggregation:

For the Pharmacy claims, we took the average of each of the ADE indicators for a particular patient (signified by a unique "therapy_id") given in the data. Since, each of the ADE indicators was a nominal categorical variable, therefore taking an average signifies the proportion of the number of times that the patient had filed Pharmacy claims for a drug that is used to treat that ADE. We treated the variables that signify whether the Pharmacy claim has any drug with a known interaction with Tagrisso ("ddi_ind"), type of claim ("clm_type"), whether the claim was for a specialty drug ("specialty_ind") or a maintenance medicine ("maint_ind") and ultimately whether the claim was reversible ("reversal_ind") and whether the prescription was filled with the mail-order pharmacy ("mail_order_ind") in the same way by taking an average. We also took an average for the prescription cost of the Pharmacy claim (signified by the variable "rx_cost") and the maximum value for the cumulative cost amount for the prescription (signified by the variable "tot_drug_cost_accum_amt").

For the Medical claims, we took the maximum value for any of the ADE indicator variables showing side effects and the same treatment was done to the primary diagnosis variable to aggregate the data at a particular patient level. This was done to convey the information whether these patients have ever been diagnosed with a side effect for their Medical claim. Before finding the maximum value, for the primary diagnosis codes, we created dummy variables for the different primary diagnosis codes for the ICD-10 format. We grouped the diagnosis into broader categories from A to Z for the longer codes given in the data. The ICD-10 codes were researched from external sources.[5]

So, to summarise we looked at the proportion of ADE events for which Pharmacy claims were filed and the occurrence of ADE events for which Medical claims were filed. Then we joined the Pharmacy claims and Medical claims data with the Target variable on the "therapy_id" field. A Left Join was performed to make sure that we do not lose the information of those patients who are not present in these two claims data sets but are present in the Target dataset. The handling of those values is described above in the section on "*Data Imputation*".

5. Exploratory Data Analysis:

This exploratory data analysis delves into some notable critical factors influencing therapy discontinuation. Specifically, we explore:

- *Side Effects and Discontinuation:* The prevalence and impact of side effects, including the influence of pre-existing conditions before therapy initiation.

  Before we start with the EDA, it is crucial to highlight that in therapies where medical claim data are missing, it is assumed that there are no diagnoses of side effects. This assumption plays a fundamental role in the subsequent analysis and insights that are derived from the data.

  Based on the chart on the right, approximately 60% of patients were diagnosed with at least one side effect, indicating that a significant majority of the patient population is contending with additional challenges during their therapy. Drilling deeper into specific side effects, fatigue is the most common side effect with 38% of diagnoses, followed by nausea diagnoses which is 20%.

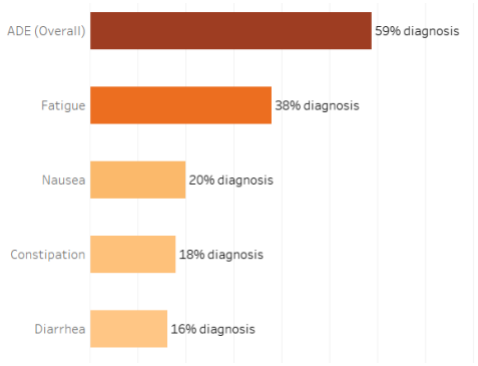


Figure 7: Prevalence of Side Effects

  For the bar chart on the left we see that 33% of patients who were diagnosed with side effects opted to discontinue their therapy, presenting a stark contrast to those who did not experience side effects and exhibited a 99% therapy continuation rate. This significant divergence underscores the profound impact that experiencing side effects has on a patient's decision to cease their therapy. The data highlights an area that may benefit from targeted interventions and support to potentially reduce therapy discontinuation rates.

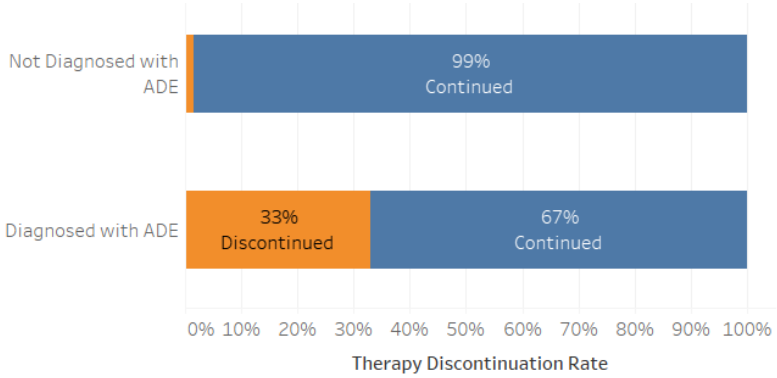


Figure 8: Impact of Side Effects on Discontinuation

  From the table below we can see that 38% of the patients have been diagnosed with Fatigue, where we see a 36% therapy discontinuation rate. Nausea, experienced by 20% of patients, is even more impactful, being associated with a 44% discontinuation rate. Constipation and diarrhea, diagnosed in 18% and 16% of patients respectively, also present substantial barriers to therapy continuation, with discontinuation rates of 39% and 34%. Notably, for patients who did not experience these specific side effects, the discontinuation rate remained consistently low, under 10%, underscoring the substantial impact that these side effects have

on therapy adherence. This is particularly noteworthy as it suggests that interventions or alternative strategies to mitigate these side effects could have a widespread impact, potentially enhancing therapy adherence for a substantial portion of patients.

Side Effect	Diagnosis Rate	Therapy Discontinuation Rate if Side Effect is Diagnosed (Side Effect is NOT Diagnosed)
ADE (Overall)	59%	33% (1%)
Fatigue	38%	36% (4%)
Nausea	20%	44% (6%)
Constipation	18%	39% (7%)
Diarrhea	16%	34% (8%)

Table 3: Side Effects Diagnosis

The bar chart on the right showcases a mere 6% discontinuation rate for patients with no side effects diagnosed either before or after the Tagrisso treatment. A significantly higher discontinuation rate of 31% is noted for those who were not diagnosed with illnesses corresponding to the side effects before treatment but experienced them afterwards. In contrast, patients who were diagnosed with illnesses corresponding to the side effects before treatment but did not experience them afterwards exhibit an 11% discontinuation rate. Strikingly, the highest discontinuation rate of 46% is seen in patients who were diagnosed with illnesses corresponding to the side effects both before and after treatment, indicating a vital need for intervention. Additionally, a 31% discontinuation rate among those developing side effects only after treatment underscores the drug's impact and the importance of managing both pre-existing and emergent side effects to enhance treatment adherence and overall outcomes. Balancing the management of both aspects is crucial to maintaining therapy continuity and improving patient experiences.

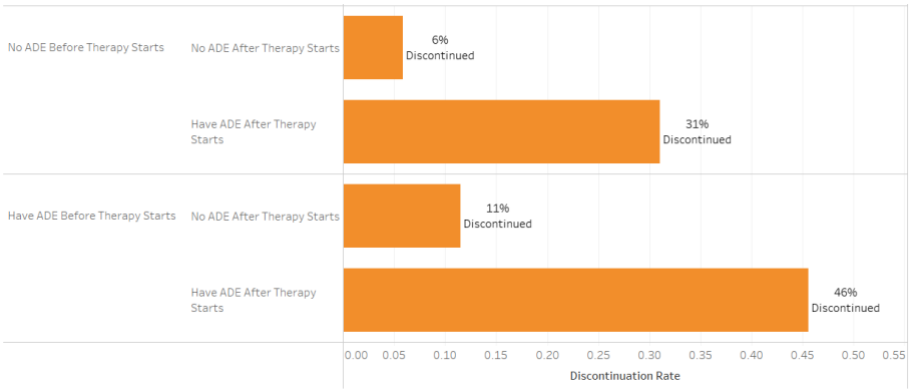


Figure 9: Pre-Existing Side Effects Prior to Therapy Initiation

- *Common Side Effect Management Drugs:* A list of the top 9 drugs frequently used for side effect management.

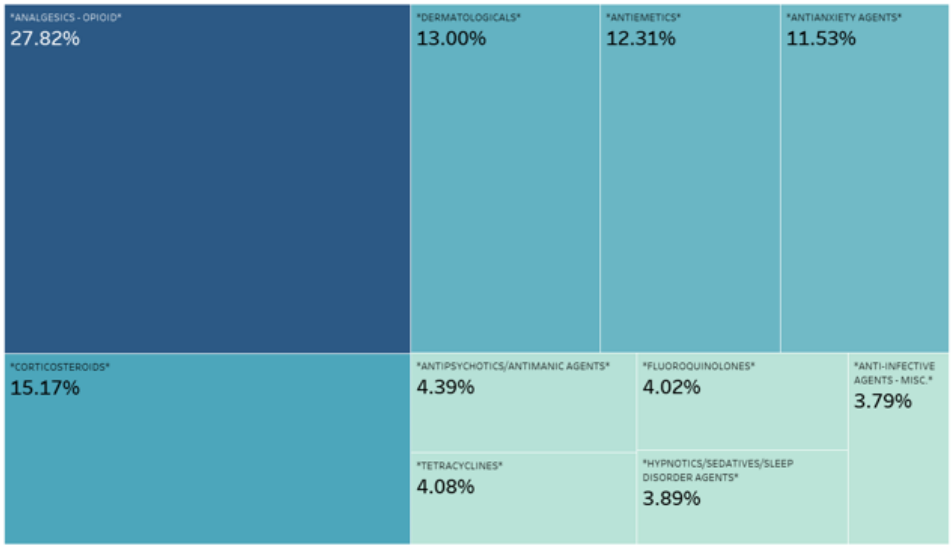


*Figure 10: Top Non-Maintenance Drugs used most Frequently for side-effects**

Grouping and Insights:

- Most Significant Proportion (>20%): Analgesic Opioids
- Moderate Proportion (10%-20%): Corticosteroids, Dermatological, Antiemetics, and Antianxiety Agents
- Smaller Proportion (<10%): Antipsychotics/Antimanic Agents, Tetracycline, Fluoroquinolones, Hypnotics/Sedative/Sleep Disorder Agents, Anti-infective Agents

Analgesic opioids, which are primarily utilized for pain management, have the most substantial proportion of usage among patients, indicating a prevalent need for pain management strategies during treatment. This could be attributed to pain being a common side effect or symptom in various conditions and treatments. The moderate proportion group ranges from corticosteroids to antianxiety agents, suggests a balanced need for managing various aspects like inflammation, skin conditions, nausea, and mental health during treatment. It indicates that a multifaceted approach to managing side effects, addressing both physical and mental aspects, is crucial. The smaller proportion group, which includes antibiotics, sleep disorder agents, and anti-infective agents, might indicate specific, but less common needs among the patient population. Additionally, it is pivotal to note that while these drugs play a crucial role in managing and mitigating side effects, they can also introduce additional side effects, necessitating a careful and well-monitored approach to ensure that the benefits outweigh the potential risks, and that patient well-being is maintained throughout treatment.

*Check Table 1 in the appendix to get a summary of the drugs along with the side effects treatment.

- *Critical Discontinuation Period:* We identify the crucial 60–120-day timeframe after therapy initiation for therapy discontinuation.

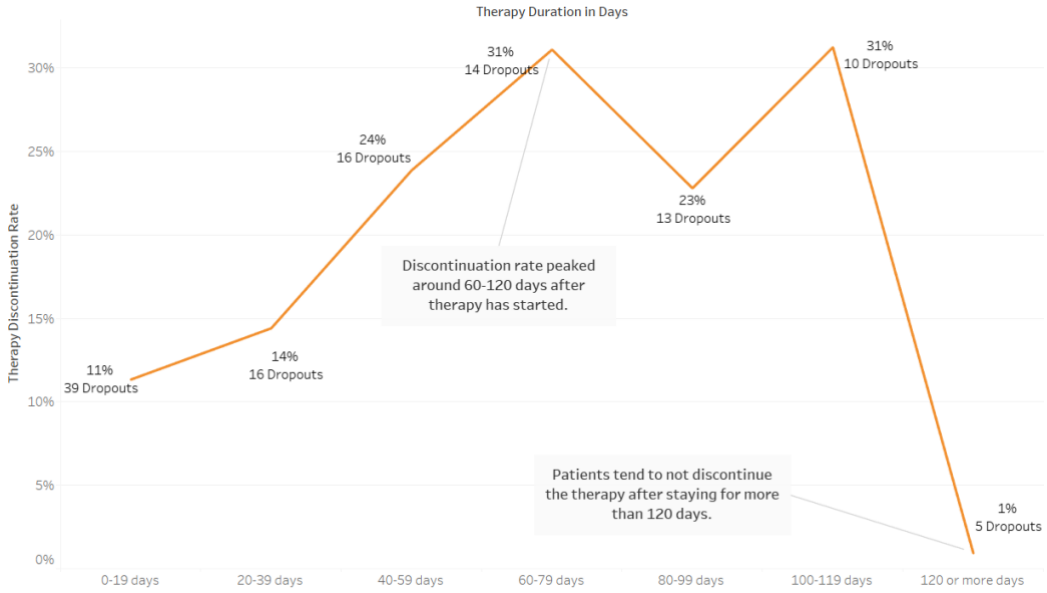


Figure 11: Critical Points for Therapy Discontinuation

The line chart, illustrates the relationship between therapy duration and discontinuation rate. The discontinuation rate experiences a significant surge and peaks within the 60-120 days range, registering discontinuation rates up to 31%. This suggests that this time frame may represent a critical period during which patients are particularly prone to discontinuing therapy. Conversely, post the 120-day mark, the discontinuation rate plummets, nearing 0, indicating a stark reduction in the likelihood of patients ceasing their therapy beyond this point.

- *Low-Income Groups and Drug Costs:* An investigation into the financial burden faced by low-income individuals when purchasing medications.

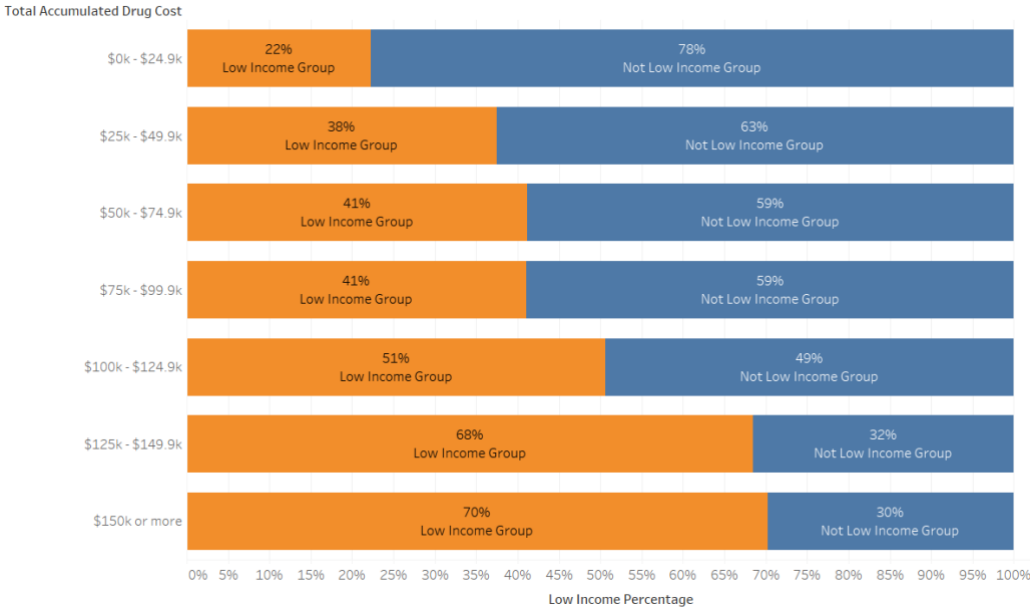


Figure 12: Drug Cost Burden for Low Income Groups

The percentage bar chart delineates a striking and somewhat counterintuitive trend regarding

the accumulated drug costs and income levels of patients. Specifically, within the lower accumulated drug cost bracket of 0-24.9k, only 22% of patients hail from the low-income group. Astonishingly, this proportion escalates significantly in the higher cost bracket of 150k or more, where 70% of the patients are from the low-income group. This paradoxical scenario, where those from the low-income group incur higher accumulated drug costs, unveils a critical socioeconomic disparity in the financial burden of healthcare. It suggests that individuals within the low-income group are disproportionately shouldering higher drug costs, which could potentially be attributed to various factors such as the lack of access to insurance, subsidized healthcare, or alternative financial aid that could otherwise alleviate some of their financial burdens.

- *Claim Submission and Discontinuation:* The link between therapy discontinuation among low-income patients and increased healthcare claim submissions.

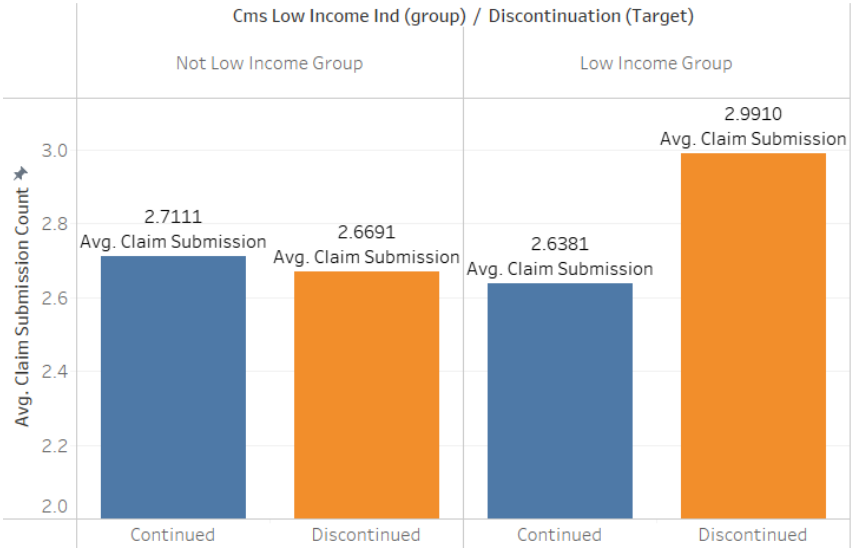


Figure 13: Discontinued Low Income Patients and Claims Submission count

The bar chart below reveals a distinctive trend in claim submissions among various patient categories, particularly highlighting a pronounced disparity for low-income patients who have discontinued their treatment. The discontinued patient group demonstrates a notably higher average claim submission that is close to 3, which starkly contrasts with the other categories: patients not in the low-income group (both continued and discontinued) and patients in the low-income group who continued their treatment. These latter categories exhibit a significantly lower rate of claim submissions compared to the category of low-income patients who have discontinued.

In the realm of claim resubmissions, it is pivotal to recognize that low-income groups are notably susceptible to claim denials. A report from Abelson (2023) [6] elucidates that private health insurance companies, paid by Medicaid, have denied a multitude of requests for care for low-income Americans, often with minimal oversight from federal and state authorities. This denial of medical care for the poor occurs at significantly high rates, with some

Medicaid plans denying medical care under requests for prior authorization of services by rates that were greater than 25% in 2019.

Moreover, understanding the reasons behind claim denials is crucial. According to a survey conducted by Experian Health, reported by Cass (2023) [7], the top three reasons for claim denials, when selected, were: Authorizations (48%), Provider eligibility (42%), and Code inaccuracies (42%). Other notable reasons include incorrect modifiers, failure to meet submission deadlines, and patient information inaccuracy.

6. Feature Engineering:

As we prepare the data for the predictive modeling phase of the problem, we have carried out a few feature engineering/pre-processing steps to introduce new features and create a dataset suitable for modeling. After these preprocessing steps, we were left with around 55 predictor variables for our modeling phase.

- *Derived Metrics:* Based on our understanding of the problem statement and exploratory analysis observations, we created a few derived metrics to capture patterns related to the physician/hospital visits made by the patient and the medications being purchased by the patients. These metrics will be leveraged in both supervised predictive modeling and unsupervised patient clustering in the later sections.
    - (Med Claims) Number of Days Between Latest Visit Date and Therapy Start Date: To capture how recently a patient had to visit the doctor for any reason since their therapy started.
    - (Med Claims) Total Count of Visit Dates: To capture how many times the patient had to visit the doctor for any reason. Our hypothesis is that more visits could potentially indicate that the patient is having more problems.
    - (Med Claims) Max of ADE Indicator and All Individual Ade Indicators (Diarrhea, Nausea, etc.): Indicates if the patient had experienced any kind of ADE at any point in time as a binary variable.
    - (Med Claims) Average Count of Medical Claims Submitted: Indicates if the patient had to submit multiple medical claims for each processed claim on average. This is calculated based on the assumption that duplicate entries in the Med Claims dataset indicate multiple claims being filed by the patient. Higher average claims could indicate financial difficulties, frustration, and higher chance of leaving.
    - (Rx Claims) Number of Days b/w Latest Process Date for EGFR Therapies and Therapy Start Date: The latest process date for EGFR Therapies (Drug Class Desc = "ANTINEOPLASTIC - EGFR INHIBITORS") indicates their most recent purchase of Tagrisso or other medicines like Tagrisso. Our hypothesis is that a larger number of days

between therapy start date and the latest purchase of Tagrisso could indicate higher likelihood for sticking to the treatment and vice versa.

- (Rx Claims) Total Drug Cost Accumulated Amt after Therapy Start Date: Cumulative cost of medicines prescribed to the patient since the therapy started. Higher costs could indicate more problems for lower income groups.
- (Rx Claims) Frequency of EGFR Therapy Claims: Lower frequency could indicate patients just starting out and higher frequencies indicate patients that have been there for longer. As seen in the EDA, different range of frequencies could have different chances of leaving.
- (Rx Claims) Frequency of Common Side-Effect Related Rx Claims: The frequency of common side effect related medications (at a patient level) like Antiemetics, Corticosteroids, Analgesics were filtered. Higher frequency for some of these medications could be associated with a higher chance of leaving treatment.

- *One-Hot Encoding Categorical Variables:* We have created dummy encodings of some key categorical variables like the First Digit of the Primary Diagnosis Indicator and Place of Treatment.
- *Removing Variables with Low Variance*: Variables having lower than 5% variance were removed from the final modeling dataset since these variables would have very homogenous information and would not be of much help in the predictive modeling process.
- *Handling Class Imbalance with SMOTE Oversampling:* As mentioned in the initial exploratory data analysis, there is a high amount of imbalance in the distribution of the target variable ("tgt_ade_dc_ind"). To handle this class imbalance problem, we decided to experiment with two methods –
  - SMOTE Oversampling: We used the SMOTE method to generate synthetic samples for the minority class and create an equal representation of the two classes.
  - Cost Sensitive Learning: In this approach, we attempted to handle the class imbalance by assigning higher weights to the cost associated with the minority class while training the models. This is inherently available as a hyperparameter in most python implementations.

Out of the two methods mentioned above, SMOTE provided slightly better predictive performance based on the AUC scores and hence we finalized the SMOTE approach for dealing with the class imbalance problem.

7. Predictive Modeling:

A. Model Choices:

Due to the large class imbalance, large number of features, and intricate relationships among the variables, tree-based ensemble models were our primary choice of models for experimentation.

These models tend to be very robust at handling class imbalances and avoiding overfitting. We decided to experiment with three models – *Random Forest,* *Light Gradient Boosting Model (LGBM),* and *CatBoost.*

Random Forest was chosen due to its robustness in handling overfitting. It is also a much simpler and easier model to tune than the gradient boosting models.

LGBM was our first choice of gradient boosting models because it was the fastest among all gradient boosting models which allowed to perform quick experimentation. Gradient Boosting Models in general were our first choice of models due to their robustness in handling imbalanced data and ability to handle complex relationships in tabular data.

Finally, we also experimented with CatBoost (another type of gradient boosting model) due to the large number of categorial features among the predictor variables. CatBoost provides the ability to handle categorical features with special encoding capabilities like target encoding without having to explicitly perform any preprocessing for these categorical variables.

We also experimented with a *Meta-Ensemble Stacking Approach* to benefit from the pros of each of the above-mentioned three models. The idea was to counterbalance the weak areas of predictive ability that one model might have with the predictive ability of the other two models. For this, we used the probability predictions of the first layer of base models (Random Forest, LGBM, and CatBoost) as predictor variables and trained a Logistic Regression Model with these new set of 3 predictor variables and the same target variable ("tgt_ade_dc_ind"). This model would allow us to combine the performance of the three models and leverage the strong areas of each model.

B. Model Selection using a 10-Fold Stratified Cross Validation:

Due to the small dataset size (1232 Unique Therapy IDs), it was important to ensure that we had a robust and reliable validation set. Hence, instead of a single train-test split, we decided to follow a Stratified K-Fold Cross Validation method for building predictive classification models. We decided to go with a larger value of K due to the small dataset size and used a 10-Fold Cross Validation stratified based on the target variable ("tgt_ade_dc_ind"). This allowed us to ensure that the models we were building had stable performance on varying subsets of the data.

For tuning the hyperparameters of the model, a mix of *Manual Tuning* and *Grid Search* approach was utilized. The hyperparameters were first tuned manually for rough approximation to reduce the hyperparameter space quickly. This was followed by a grid search finetuning find the optimum hyperparameters that provid the best performance.

C. Model Evaluation, Disparity Evaluation & Final Model Selection:

The performance metrics for all the models we have experimented with are shown in Table 4. The final model was chosen based primarily on the 10-Fold Cross Validated AUC Score shown in the table above. In addition to that, as a secondary measure, the complexity of the model was also

considered. Hence, the final model was chosen based on a tradeoff between these performance metrics and complexity factors.

#	Model Name	Model Hyperparameters	10-Fold Cross Validated Avg. AUC Score
1	Random Forest	Number of Estimators: 300 Maximum Depth: 60 Minimum Weight Fraction Leaf: 0.0001	0.9548
2	**Light Gradient Boosting**	**Learning Rate: 0.000001** **Feature Fraction: 0.5** **Bagging Frequency: 5** **Bagging Fraction: 0.3** **Number of Leaves: 10** **Minimum Data in Leaf: 10** **Maximum Depth: 7** **Minimum Sum Hessian in Leaf: 0.0001**	**0.9659**
3	CatBoost	Learning Rate: 0.01 L2 Leaf Regularization: 5 Subsample: 0.5 Max Depth: 5 Minimum Data in Leaf: 12	0.9611
4	Stacked Ensemble	Logistic Regression Parameters: Penalty: L2 Regularization Strength (C): 0.0001 Solver: LBFGS	0.9681

Table 4: Model Building and Performance

For our final model, even though the Stacked Ensemble model performed the best out of all the models, we decided to go with the Light Gradient Boosting Model since the performance gain from the stacked model were not large enough to warrant the large increase in complexity of the model and also the additional computational overhead that it would add if the model were to be deployed at any point in the future. It would also give us better interpretability for understanding how the model is making decisions.

We also analyzed the disparity score of the final LGBM model based on the *Adverse Impact Ratio (AIR)*. It is a method that has been traditionally used for analyzing employment discrimination and lending biases [8]. We chose the metric due to its simplicity of interpretation. It is calculated as the ratio between the Selection Rate in Minority Group and the Selection Rate in Majority Group. The selection rate in our case will be the rate of therapy discontinuation. AIR values anything above 0.8 is acceptable and unbiased [9]. Figure 14 shows the AIR value and the associated p-values for the LGBM Model. We can see that all the AIR values are above 1 except for the Asian Race Category. However, the p-value for this indicates that the result is not significant enough to make a conclusion

about the presence of bias. Hence, we can safely assume that our model is unbiased at an overall level.

Adverse Impact Ratio Summary Table

* Percent Missing: Race: 0.00%, Sex: 0.00%

Group	Reference Group	Group Category	Total	Favorable	Percent Favorable	Percent Difference Favorable	AIR	P-Values	Practically Significant	Shortfall
Asian	White	Race	152.0	11.0	7.24%	5.92%	0.550	0.054	No	
Black	White	Race	145.0	25.0	17.24%	-4.08%	1.310	0.190	No	
Hispanic	White	Race	42.0	8.0	19.05%	-5.89%	1.447	0.252	No	
Native American	White	Race	5.0	1.0	20.00%	-6.84%	1.520	0.509	No	
Other	White	Race	37.0	7.0	18.92%	-5.76%	1.437	0.321	No	
White		Race	699.0	92.0	13.16%					
Female	Male	Sex	815.0	112.0	13.74%	-2.66%	1.241	0.246	No	
Male		Sex	334.0	37.0	11.08%					

Figure 14: Adverse Impact Ratio Calculations

Interpretation of Model Outcomes using SHAP Values:

To understand how the model is making decisions and analyze what makes a person leave the treatment, we analyzed the Top 20 features from the Light GBM Model using SHAP Values. This helped us understand the impact of a variable on the model outcome and the direction of that impact. Figure 15 below shows the SHAP values and the direction of the impact for the top 20 variables.

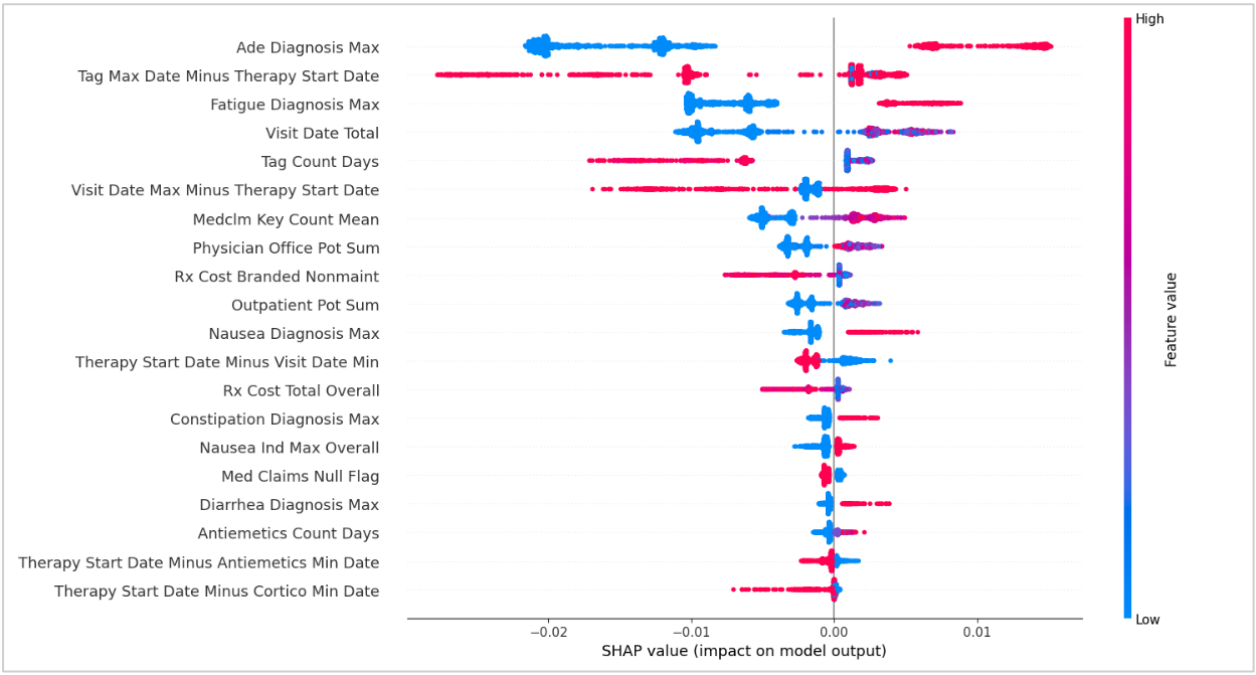


Figure 15: SHAP values from the LightGBM model

A few key observations that we made based on this analysis were as follows:

- *Med Claims* variables:
    - Lower occurrence of Ade Diagnosis (Ade Diagnosis Max) indicates a higher chance of person not leaving the treatment which confirms the fact that the occurrence of side-effects is indeed a major factor in people leaving the treatment. This variable is also the most important variable for the model indicating a strong relationship between side-effect occurrences and people leaving the treatment.

- Fatigue and Nausea (Nausea Diagnosis Max and Fatigue Diagnosis Max) are the two most important side-effects based on the model that seem to dictate whether someone is likely to leave or not. Lower occurrence of both these side effects can indicate that a patient is less likely to leave.
- A larger number of visits to the Physician or as an Outpatient at a hospital (Physician Office Pot Sum & Outpatient Pot Sum) indicates that a patient could have had problems that required them to visit the doctor multiple times and hence a lower number of visits indicates a lower chance of leaving the treatment.
- As we observed in our initial EDA, patients having to file a larger number of claims (Med Claims Key Counts) before getting their claim processed are also more likely to leave the treatment. This could indicate frustration.

- *RX Claims* Variables:
  - The larger the duration between the Therapy Start Date and the Latest Date of Tagrisso (Tag Max Date Minus Therapy Start Date) purchase, the higher the chance that the patient is not going to leave the treatment. It indicates that patients sticking around for longer beyond a certain point are less likely to leave despite experiencing ADEs. This ends up being the second most important variable for the model.
  - Higher cost for Tagrisso (Tag Spending) indicates that the patient is religiously continuing their treatment plan and hence indicates a lower chance of a patient leaving.
  - The Nausea Medication Indicator (Nausea Ind Max Overall) which indicates the purchase of nausea treatment medications also directionally aligns with the impact of Nausea Diagnosis indicator from the Med Claims data. This corroborates the fact that nausea is a highly occurring problem which can lead to patients leaving.

To understand the model decision-making process in a much more comprehensive manner, we followed up the global feature importance and contribution analysis with a local interpretation analysis. For this, we randomly picked a few observations from both the target classes, calculated SHAP values for the observations, and plotted Force Plots to understand the driving factors behind an observation being classified as leaving treatment and not leaving treatment. The red features indicate the factors pushing the prediction closer to Target = 1 (discontinuation) and the blue features indicate the factors that push the prediction towards Target = 0 (therapy continuation)

- *Sample 1 (Target = 1):* Figure 16 shows the force plot for a patient that has left the treatment. We can see that the patient having experienced both nausea and fatigue within just 1 month of starting Tagrisso (tag_max_date_minus_therapy_start_date = 28) has pushed the predictions towards Target = 1 with a very high probability, i.e., the patient is highly likely to discontinue treatment.

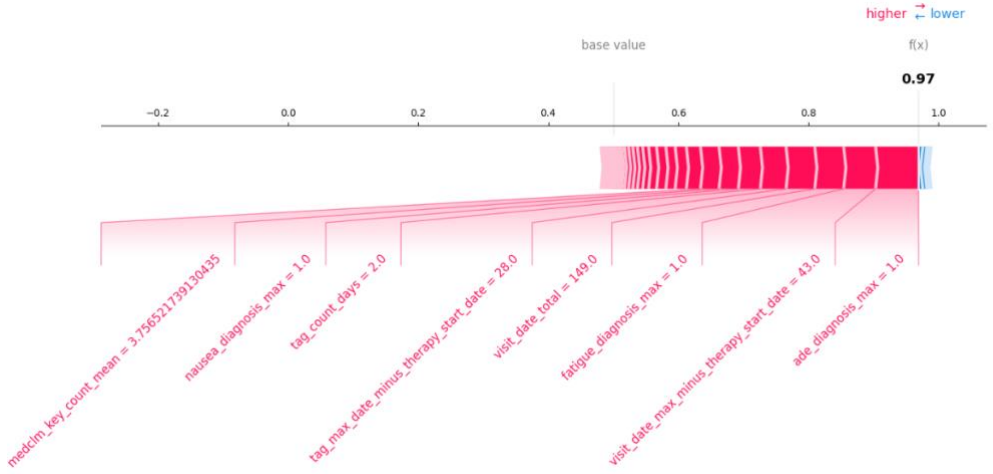


Figure 16: Second Example of Local SHAP values to explain the Predictive Model

- *Sample 2 (Target = 1):* For this example, we can see that the patient not having experienced any side effect nudges the prediction towards 0. However, this is counterbalanced by the fact that the patient has purchased antiemetics and has had to visit the physician multiple times. The patient also seems to have to file multiple claims submissions on average before getting it processed (medlcm_key_count_mean = 3.2, 3 claims filed on average before being processed). They have been on Tagrisso for about 120 days and the predictions are being nudged towards Target = 1 due to a combination of claims processing frustrations and common side effects like nausea, which might have not just been logged by the physician as a potential ADE.

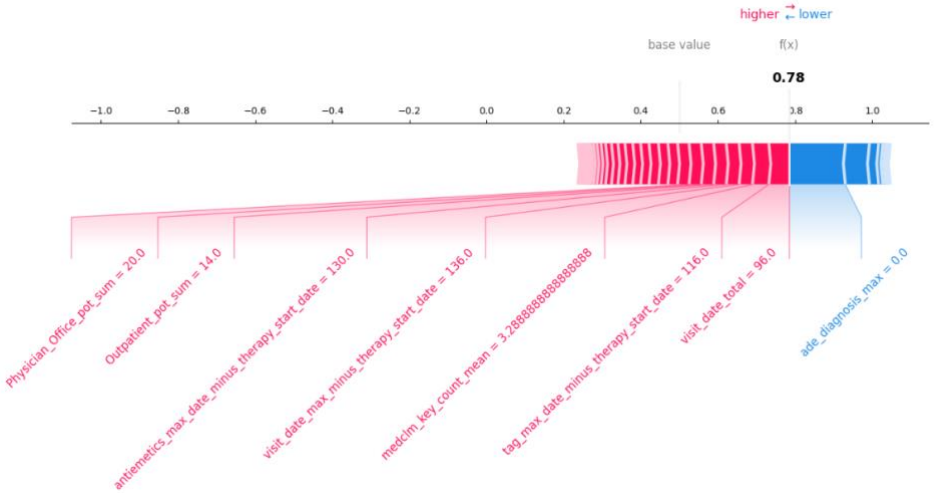


Figure 17: Second Example of Local SHAP values to explain the Predictive Model

- *Sample 3 (Target = 0):* Based on the force plot (Figure 18), this patient has not experienced any fatigue related side-effects or any other ADEs within the first 1 month of their course of Tagrisso which is pushing this patient to be tagged as most likely to be Target = 0. It might be beneficial to track and monitor early-stage patients like these to understand their progression and take proactive measures.

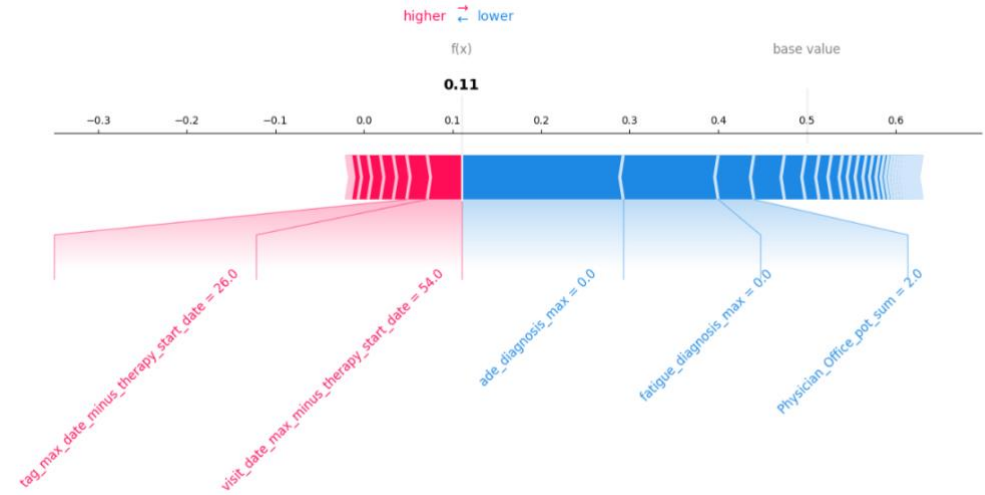


Figure 18: Third Example of Local SHAP values to explain the Predictive Model

8. Post Modeling Analysis:

In this segment, we will go through a segmentation procedure that helps us identifying the least vulnerable group through to the highest vulnerable groups. The least vulnerable group is the one on whom Humana does not need to immediately focus on as they are the least likely to leave the treatment. On the other hand, the patients in the most vulnerable should be targeted on priority as they are the most likely to leave the treatment within this 6-month period.

A. Clustering:

We used a few important variables such as the Low-Income Indicator, Disabled Indicator, Age, Indicators for different ADE diagnoses, Cost of the prescription for which a Pharmacy claim was filed, and the number of days that the patient was part of the Tagrisso medication process. Since we had both categorical and continuous variables, we adjusted the clustering procedures with the help of Gower Distance [1] and ended up creating 4 clusters.

Cluster 1 (High Risk Seniors):

There are around 15.26% patients in this category and majority of them do not belong to the low-income class and have an average age of 75.17. But we see that on an average they have filed highest number of Medical claims for the nausea and fatigue diagnosis. In general, these patients have some of the highest proportion of ADE diagnosis for the different claims. Also, these patients seem to have a lower cost for the Pharmacy claims (around $127.14) but their inpatient (3.28 on an average) and ER visits are some of the highest amongst all the patients.

Cluster 2 (Vulnerable & Economically Challenged):

There are around 10.34% of the patients in this cluster (around 99.9%) belong to the low-income category. We can see that the patients in this category have a lot more side effects on an average (around 4.13) and therefore these people are more likely to leave the treatment. The proportion of people who are disabled in this category is around 33% and with a very high number of Medical claims filed for both nausea (around 4.67 on average) and fatigue (around 7.10 on average)

Cluster 3 (Pre-Dropout Patients):

26.89% patients belong to this cluster with very less people (less than 3.5%) leaving the treatment. But even in this group, we see that proportion of people filing for Medical claims regarding nausea is around 0.76. We also notice that majority of the people (around 89.5%) receive low-income subsidies from the CMS but has a comparatively high (around 23%) disability percentage.

Cluster 4 (Persistent Patients):

This category has the greatest number of patients (around 47.5%) with the least number of patients leaving the treatment. We notice that the proportion of people suffering from various ADE side effects is also the least for this group. Subsequently, we notice that the amount of money for which these patients have filed for Medical and Pharmacy claims are some of the lowest.

The table below summarizes the different variables based on which the clusters are defined and segregated from one another.

Variables	*Cluster 1 (High Risk Seniors)*	*Cluster 2 (Vulnerable and Economically Challenged)*	*Cluster 3 (Pre-Dropout Patients)*	*Cluster 4 (Persistent Patients)*
Percentage of Patients Leaving the treatment	33%	29%	3%	1%
Average Age	75.17	73.92	72.92	73.74
Economic and Disability Status	Highest income group with less proportion of disability	Very low income and highest proportion of disability	Low income and moderately high proportion of disability	High income and low proportion of disability
Nausea and Fatigue Diagnosis	Highest number of claims filed for Nausea and Fatigue.	Second highest number of claims filed for Nausea and Fatigue.	Very small number of claims filed for Nausea and Fatigue.	Lowest number of claims filed for Nausea and Fatigue.
Proportion of people suffering from ADE	Very high	Highest	Low	Lowest

Comparison of Pharmacy Claims to Medical Claims	Pharmacy Claims for very low amount of money were filed but very expensive Medical claims were filed.	Low number of Pharmacy claims but one of highest number of Medical claims.	Both Pharmacy and Medical claims were filed for less amount of money.	Both Pharmacy and Medical claims were filed for the least amount of money amongst the 4 clusters.

Table 5: Cluster Summary

B. Survival Analysis at Segment Level:

To corroborate the cluster analysis, we followed it up with Survival Analysis to estimate the time-to-event (the event being therapy discontinuation) for each of the segments. It is especially effective in dealing with censored data, where not all patients have experienced the event of interest (therapy discontinuation in our case) by the end of the analysis period. [11] By analyzing the survival probabilities (chance of continuing therapy) over time, we can obtain valuable insights into the differences in survival rates among patient clusters. These survival probabilities over time will help us understand the most vulnerable point in time since the therapy start dates for each of these patient clusters. This information can in turn be leveraged to identify the optimal intervention points for each patient cluster. Finally, the rate of decrease in survival will also help us validate whether the vulnerability definitions that we have assigned for our clusters are accurate.

Data Preparation and Modeling Process:

For the survival analysis study, since our goal was simply to understand if the rate of survival differed across the different patient clusters, we simplified the problem and used a univariate Kaplan-Mier Estimator survival analysis model to estimate the survival functions for each cluster. [12] The Kaplan-Mier estimator takes in just two inputs – duration and event occurrence indicator. The data for the survival analysis model was prepared by calculating the duration between the Most Recent Process Date in the Rx Claims dataset and the Therapy Start Date. The original target variable (tgt_ade_dc_ind) was used as the event occurrence indicator. We built a separate survival model for each cluster and estimated different survival curves for each of these clusters. The survival curves are shown in Figure 19.

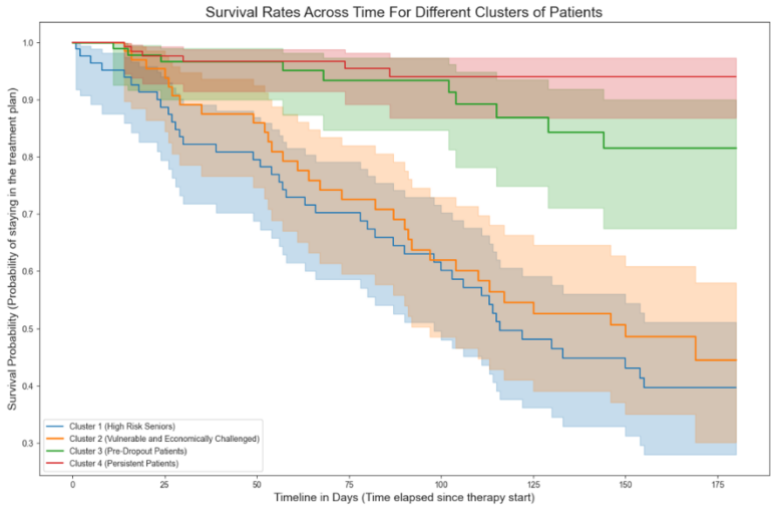


Figure 19: Graphical Representation of Survival Analysis

Interpretation:

Figure 18 shows us that the survival curves obtained from the Kaplan-Mier model align with our vulnerability classification from the patient clustering process.

Cluster 1 or the High-Risk Seniors have a lower chance of survival right from the get-go and has the lowest survival probabilities among all the clusters.

Cluster 2 or Vulnerable and Economically Challenged group has the second lowest survival probabilities. They start off well with higher survival probabilities compared to Cluster 1 but start seeing a drop in survival probabilities around the 25 to 30-day mark. Another sharp drop for both Cluster 1 and Cluster 2 happens around the 50 to 70-days range. This indicates a need for early focus on these patient clusters and understand their problems right from the beginning to be able to help them continue their treatment plans.

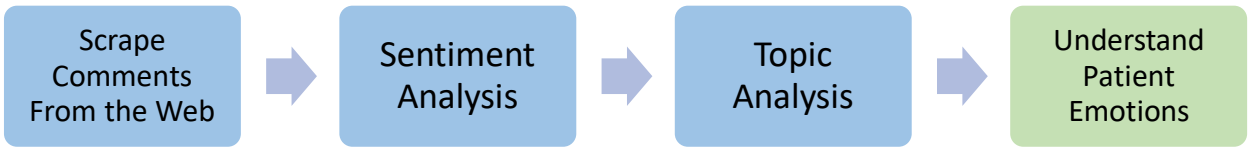
Cluster 3 or the Pre-Dropout Patients seem to have high survival probabilities in the first 100 days and follows Cluster 4 (the least vulnerable group) very closely. However, a sharp drop starts happening around the 100 to 120-day range and it keeps dropping from then on. This could be due to financial implications since there is a significant proportion of low-income patients in this group or could be because of a slower manifestation of side effects. It indicates a need for continuously tracking these patients and understanding the reason behind the build-up of their problems before they reach the tipping point so that these patients do not end up discontinuing the therapy.

Cluster 4 or the Persistent Patients have the highest survival probabilities throughout the timeline and these patients do not seem to indicate a major risk of leaving the treatment throughout the timeline which aligns well with our cluster analysis and vulnerability assessments. The flat line after the 80 to 90-day mark for this group also indicates a large interval and that some of these patients could potentially be newer patients that have been on the treatment for a shorter period. It is important to passively keep in touch with these patients for proactive resolution of their problems when they start to pop up.

C. Sentiment Analysis and Topic Analysis of Online Patient Comments/Reviews

Understanding patient sentiments and experiences related to medical treatments is crucial for enhancing healthcare services and treatment outcomes. The cancer journey is a long one and patient perceptions can change over time. Hence, continuously studying patient perceptions about various aspects of the entire treatment process can help in optimizing the treatment experience and identifying pain points. This can help increase retention and patient satisfaction. [13] [14] Hence, to facilitate this, we have followed two Generative AI-driven NLP approaches for understanding patient sentiments with regards to Tagrisso. These two approaches are:

- *Sentiment Analysis:* To understand the emotional journey that the patients undergoing Tagrisso treatment are going through.
- *Topic Analysis:* To find out what patients are mainly talking about and identify the key breaking points in the entire patient journey.



We started this process by scraping publicly available data on Tagrisso reviews and patient experiences from three medical/drug related websites - *Drugs.com, WebMD, and HealthUnlocked*. These websites contained reviews about Tagrisso and patients sharing their Tagrisso treatment experiences anonymously. We were able to obtain about 200 reviews through the scraping process. Post the scraping process, the data collected was cleaned, and Microsoft Azure Open AI GPT3.5 Model was used to perform sentiment analysis on the comments using prompt engineering. The comments were classified into one of the following sentiments: *Fear, Grief, Anxiety, Frustration, Hope, Relief, Loneliness/Isolation, and Inquisitive.* Once the comments were classified according to the sentiments, the top 3 sentiments were filtered, and Open AI Ada Embedding Model was used to generate embeddings for the comments. K-Means clustering was applied on the embeddings to group the comments within each sentiment into 2-3 semantically similar groups. These groups were then passed on to Open AI GPT3.5 Model for extracting the keywords/key topics in each group. Figure 20 represents the process used for understanding the patient's sentiments.

Approximately 200 reviews/comments that we gathered were analyzed and Figure 20 shows the key sentiments that were shown by the patients in these comments/reviews.

The figure shows us that Hope and Relief are the two most common positive sentiments which could indicate that the patients are going on to these platforms to share positive impacts of the medication that they are taking or just share confidence with others that are just starting out on the medication. Inquisitiveness is the second most common sentiment which indicates that patients resort to these

platforms to ask questions about the treatments, side-effects, experiences. They might be treating these platforms like an online community. Finally, the two most common negative emotions are Frustration and Fear. These emotions also make logical sense because patients could be sharing their thoughts about the side-effects they are facing, the medication not working out well for them, or general problems related to cancer treatment. While these emotions tell us about what patients are talking about on a high level, it is also important to understand the topics of discussion and dive deeper into the top emotions to understand the main drivers.

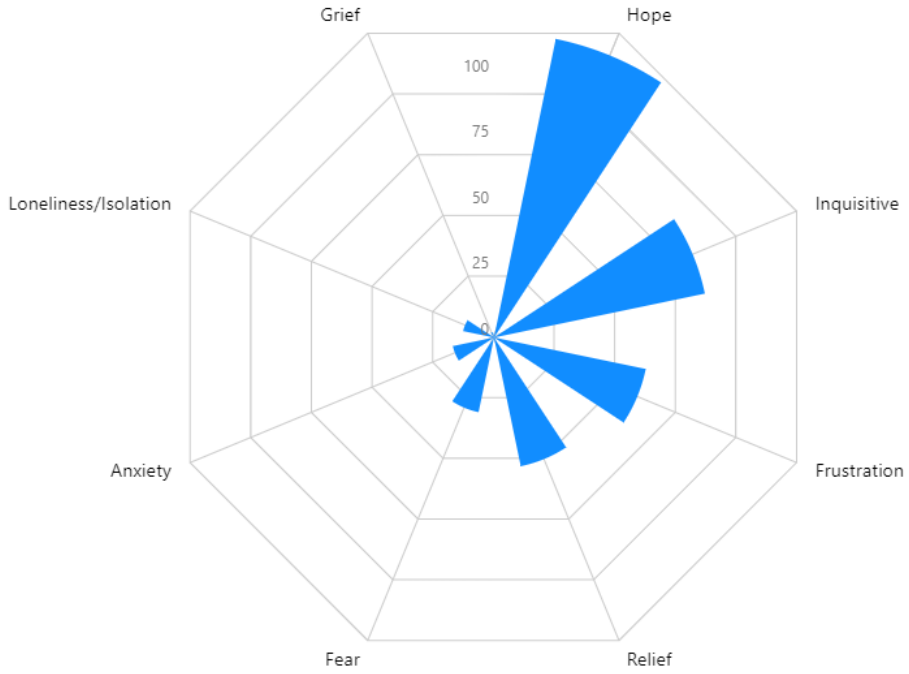


Figure 20: Radar Chart for Sentiment Analysis

In order to understand the key topics of discussion for the patients on these platforms, we have taken the top 3 sentiments and carried out topic extraction and analysis on the reviews/comments for these top 3 sentiments (Hope, Inquisitiveness, Frustration). Table 6 shows us the key topics associated with each of these 3 sentiments and a few examples.

Sentiment	Key Topics of Discussion	Examples
Hope	• Effectiveness of Tagrisso • Improving Prognosis • Manageable Side Effects like rash, mouth sores, acne, bowel issues, etc.	*"I have been on Tagrisso for 18 months. Tumor has shrunk considerably and some places no longer show up on the PET scan."* *"It does cause bowel issues so you have to be careful of that. I've also noticed an extra dryness in my mouth especially at night when I'm sleeping."*
Inquisitive	• Questions on Side-Effects	*"Are there any other drugs available with fewer side effects?"*

	• Questions about treatment options and effectiveness of Tagrisso • Seeking Emotional Support	*"Can I go to the dentist while taking Tagrisso? Are there any precautions I should take?"* *"How do I move forward and feel normal after being diagnosed with Stage IV lung cancer? Any advice from others who have been through it?"*
Frustration	• Frustration from Side Effects like nausea, digestive tract issues, acne, change in taste, etc. • Emotional/Financial challenges and frustrations	*"The cost is $5000 per month over and above their medical insurance contribution, which is not sustainable for them."* *"My mom is on Tagrisso. She is experiencing very bad nausea in the morning and early afternoon."*

Table 6: Topic Analyses

The examples highlight that patients tend to have a ton of questions on the associated side effects of Tagrisso. They also use these specific digital media platforms to share their experiences – both positive and negative. These can help us identify common pain points faced by the patients like lack of access to reliable information about side effects and treatment expectations, lack of support systems, and financial burdens for some patients.

In our topic analysis process, we also used the GPT3.5 Model to extract the side effects from the reviews/comments and created a word cloud to analyze the most mentioned side-effects by the patients. The most common side-effects mentioned by patients seem to be nausea, diarrhea, weight loss, and mouth sores. The pattern observed in the word cloud validates the observations that we made based on Humana provided data. Figure 21 shows the word cloud for the side effects.

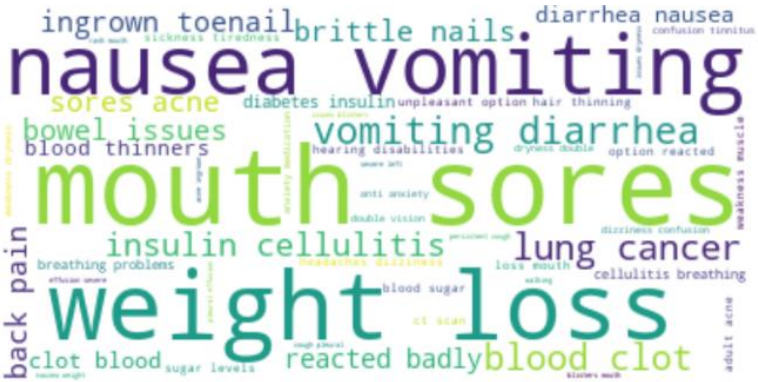


Figure 21:Word Cloud after analysing the Online Forums

9. Recommendations and Solutions:

A. Redefining the Customer Experience:

To tackle the various problems highlighted in the customer journey in the [Preliminary Research section], we recommend Humana to adopt a three-pronged approach. We have broadly categorized the issues into three segments which addresses the concerns related to medical, financial, and patient

support problems. Before delving deeper into the recommendations, we propose a new and improved 8 Step '*Illness to Wellness*' journey to ensure patients are properly heard and cared for.

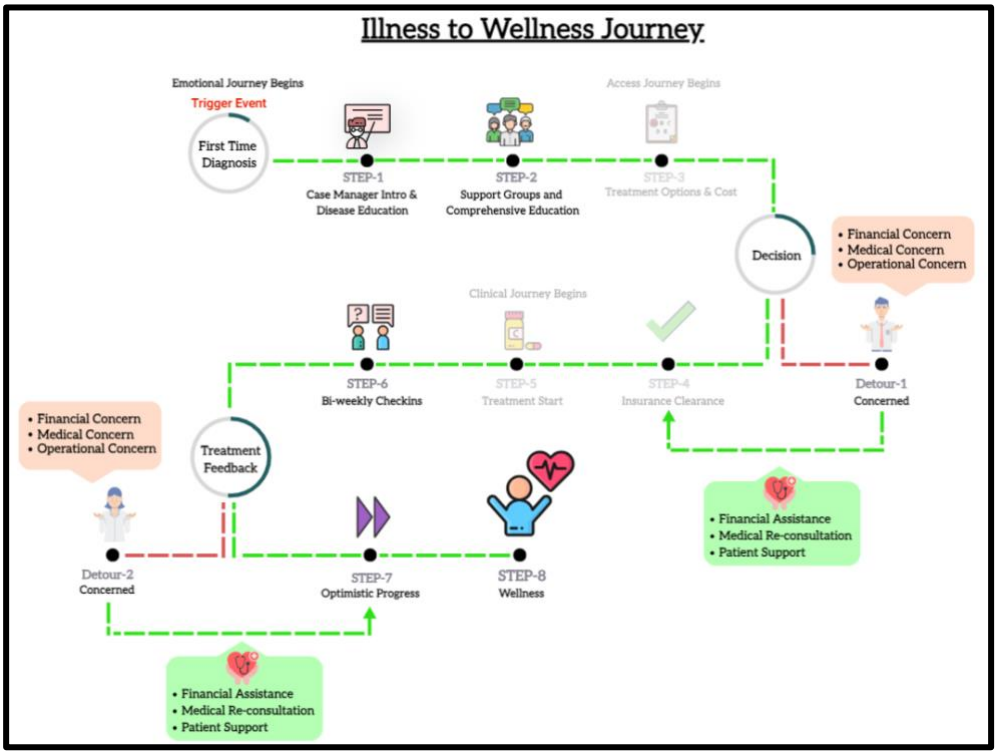


Figure 22:Illness to Wellness Journey

Our proposal considers an addition of steps where vulnerabilities were previously present. The additions are as follows:

- *Case Manager Intro & Disease Education (STEP-1):* Ensuring patients and their caregivers are thoroughly educated about the treatment journey.
- *Support Groups and Comprehensive Education (STEP-2):* Support Groups offer patients comfort through shared experiences, reducing isolation. Meanwhile, a detailed education plan covers all facets of their journey, from financial to lifestyle changes. Together, these measures provide comprehensive patient support.
- *Treatment Cost & Options (STEP-3):* This step provides transparent treatment plans (which includes appropriate costs) to build early trust and ensuring patients feel cared for.
- *Detour-1 (Concerned):* It is possible that patients still have some questions related to finances, consideration of their medical history or even general changes to their lifestyle. This is one of the two important steps on which our recommendation is based. The three-pronged approach we followed is:
  - Financial Concerns: Reduced co-insurance, deductibles, and assistance with claims processing.
  - Medical Concerns: Re-consultations through home care and virtual care.
  - Operational Concerns: Channelized patient support initiatives through case managers, support groups and comprehensive education.

- *Insurance Clearance (STEP-4):* This step is merely a checkpoint for the case manager to inform the patient about the financial details and to mentally prepare them for the journey ahead.
- *Treatment Start (STEP-5):* Onset of the patient's clinical journey.
- *Bi-weekly Check in (STEP-6):* This step creates consistent communication between the case manager and patient, ensuring accountability and enabling real-time data tracking. Bi-weekly feedback helps Humana detect patterns in side-effects and early signs of patient concerns, enabling prompt intervention.
- *Detour-2 (Concerns):* Yet again, the recommendation to solve the patient's problems in this step is like the first detour but with the addition of personalized information reports through generative AI, which case managers can leverage to understand their patients' needs better. This is the second important vulnerable point on which most of our recommendations are based.
- *Optimistic Progress (STEP-7):* Despite positive progress, case managers must stay attuned to patients' challenges, ensuring early detection and support for their well-being.
- *Wellness (STEP-8):* After recovery, patients continue attending support groups, aiding others on the same journey. Humana can share their success stories through blogs and interviews, strengthening their bond with the Humana community.

Now that our Illness to Wellness framework is set, the following recommendations will address what can be done by Humana in specific steps to ensure patient satisfaction.

B. Recommendation – 1: Financial Assistance

In this section, we explore various financial assistance programs and strategies designed to alleviate the economic challenges encountered by patients.

"What is the problem?"

- **Disparities in Drug Costs and Income Levels:** [Refer to EDA], showcases a glaring inequality in the drug costs incurred by individuals across various income strata. Notably, those in lower income brackets disproportionately shoulder significant drug expenses. A critical point in the patient journey, specifically detour-1, involves the confrontation of treatment options and their associated costs. Similarly, during the provision of treatment feedback, patients—potentially having depleted their savings might opt for discontinuation due to financial strain.
- **Differential Claim Submission Rates among Income Groups:** Detailed in [Refer to EDA], the divergence in claim submission rates, particularly the pronounced rates among low-income patients who have discontinued their therapy (averaging nearly 3 submissions), hint at a multifaceted issue. This challenge surfaces in step 4 of the Illness to Wellness journey, which necessitates claim submissions.

These revelations amplify the urgency for meticulously crafted interventions to mitigate the financial hurdles faced by patients.

"What is the solution and where do we intervene?"

I) <u>Halving Deductibles for Initial Therapy Duration</u>

a. *<u>Contextual Insight:</u>* According to Bailey (2022),[15] high-deductible health plans can influence patients to curtail their healthcare utilization, including essential preventive services, to avoid substantial healthcare expenses. Based on our findings in the EDA section, we noticed that almost 60% of the overall population experiences at least one side effect. The prevalence of common side effects paired with an average cost of $15,000 (per refill) for Tagrisso adds to the burden of a higher overall cost of medicines claimed by a patient.

b. Data extracted from the Humana website reveals an average deductible hovering around $3,000, which could pose a significant burden to financially constrained groups (Humana, 2023).

c. *<u>Utility in detour – 1:</u>* This reduction can come in if there is any financial concern expressed by the patient even before the start of the treatment. Humana must carry out background research to identify the financial vulnerabilities of the patient and understand how the therapy plan can be tailored to their needs. A 50% reduction in deductibles for the initial six months of therapy is proposed as a viable buffer against abrupt financial strain.

d. ***Utility in detour – 2:*** Similarly, we propose a 50% reduction in deductibles for any patient that expresses their financial concern for the first time after the start of the treatment.

e. *<u>Cost of Implementation:</u>*

Cost Applied to Humana for reducing Deductibles = *Average Humana's Plan Deductible* × *50%*

II) <u>Waiving Co-insurance for Initial Therapy Duration (Pays nothing after deductible)</u>

a. *<u>Contextual Insight:</u>* 20% co-insurance is the most prevalent rate incurred by patients (MetLife, 2023).[16] In the cluster analysis section, it was observed that the 'Vulnerable & Economically Challenged' group had the highest proportion of people suffering from ADEs. Further, this group also had the highest number of Medclaims among all the clusters which adds to the burden of a higher overall immediate payment. As our priority is focused towards the most vulnerable groups at first, a lower co-insurance payment would significantly help patients with their financial distress.

b. *<u>Utility in detour – 1:</u>* For the economically vulnerable, waiving their co-insurance for the initial six months of the therapy, specifically for drugs excluding Tagrisso, can potentially alleviate the immediate financial burden before the start of therapy.

c. *Utility in detour – 2:* Progressing along the same lineage of thought, if the patient expresses a financial concern at any point in time after the start of the therapy, Humana can waive co-insurance for them.

d. *Cost of Implementation:*

Cost Per Patient (6 Months) for Additional Medications Cost (Other Than Tagrisso) after Therapy Start Date = *Average Cost of RX Claims Per Patient Per Pharmacy Visit ($)* × *Average Count of Pharmacy Visits Per Patient in 6 Months* ---------- (i)

Cost Applied to Humana for 100% Reduced Coinsurance =

Cost Per Patient (6 Months) for Additional Medications Cost (Other Than Tagrisso) after Therapy Start Date × *100%* ---------- (ii)

While our recommendations suggest a 50% reduction in deductibles and waiving co-insurance at different stages of a patient journey, it is important to note that Humana can adopt our framework to change up the percentage of these financial assistance metrics however they deem necessary to arrive at a benchmark which will be beneficial for the patient and the firm. The detailed [section link] cost analysis in the sections to follow will allow Humana to get a better understanding of overall costs and their impact on the firm.

III) Streamlining and Supporting Claim Submissions

a. *Contextual Insights:* A report by Abelson (2023) [6] highlights that private health insurers, paid through Medicaid, have extensively denied care requests for low-income Americans, often without adequate oversight. Additionally, a survey by Experian Health, cited by Cass (2023), pinpoints the primary causes for claim denials as: Authorization issues (48%), Provider eligibility discrepancies (42%), and Code inaccuracies (42%). This external research observation attests our findings in the exploratory data analysis section where the low-income group had a significantly higher average number of claim submissions (3) compared to the other group hence highlighting the need for guidance in the submissions.

b. *Utility in the entire process:* Enhancing Humana's existing claim submission support through the following recommendations:

- Utilize existing digital channels to deliver educational webinars and update online claim guides.
- Enhance customer service by training existing staff for specialized claim submission assistance during specific hours.
- Institute a peer support network where experienced claimants help new patients through the submission process.
- Optimize the current claim platform by refining the user interface and adding instant automated feedback on submissions.

C. Recommendation – 2: Medical Reconsultation:

"What is the problem?"

The need for re-consultation could possibly come up in a patient's mind either looking for a second opinion for their first-time diagnosis or in general, when they have side effects. In our Illness to Wellness journey, we observe detours from the expected path to address the circumstances, wherein the patient may need additional medical advice.

- *High overall age among clusters:* As observed in the cluster analysis, the two most vulnerable clusters ('Vulnerable & Economically Challenged' and 'High Risk Seniors') has an average age of 73.92 and 75.17 years respectively. The proportion of people discontinuing therapy in these groups is 29% and 33% among its respective populations, which highlights how pervasive this issue is. The need for accessibility becomes even more important in these groups.
- *Disability status:* Cluster 2 (Vulnerable and Economically challenged) and 3 (Pre-dropout Patients) have two of the highest proportion of disabled individuals indicating issues with mobility and movement. The former group also observes the highest side effects on average, among all groups hence our recommendation must provide an easy way to obtain medical assistance for these patients, in times of need.

"What is the solution and where do we intervene?"

I) **Home Care**

a. *Contextual Insight:* Home care, also termed as Home-based primary care (HBPC), has been acknowledged for its transformative effects on patient care, especially with dormant and homebound patients. With the advancements in portable medical technology, the promise of HBPC is more tangible than ever. Considering the rapidly aging population, around 10,000 boomers are expected to join Medicare daily until 2029[1]. Distressingly, as of 2011, of the 2 million homebound individuals, only 12% reported receiving HBPC [17]. Such numbers underline the widening gap between need and service delivery, which home care is primed to bridge. Humana's Home-based primary care and preventive services is a great technology that can be leveraged in this process to address the needs of the most vulnerable clusters exhibiting prevalence of numerous old patients with disabilities.

b. *Utility in Detour-1:* For individuals facing the anguish of a first-time cancer diagnosis, the comfort of home combined with the assurance of a medical professional can be crucial. Humana's home-based primary care, in such instances, serves a dual purpose - they offer a peaceful environment for patients to adjust with their diagnosis and they facilitate the process of obtaining a second opinion. By investing in this approach, not only do patients access top-tier medical opinions, but they also receive care that underscores their priority and instills trust.

c. *Utility in Detour-2:* Adverse drug events like fatigue, nausea, constipation etc. especially from treatments like Tagrisso, can be daunting. It is imperative to provide timely medical interventions, and home care is the answer. Rather than navigating the logistical challenges and stress of a hospital visit, patients can receive immediate, personal medical attention at home, drastically enhancing their experience and potentially their outcomes.

d. *Special Consideration for Older Individuals:* As per projections, in the next 20 years, nearly half of the population aged 85 and older will require assistance with at least one activity of daily living [17]. The implications of these statistics are clear: older individuals, especially those with limited mobility or other pre-existing conditions, stand to gain enormously from home care. The elimination of hospital visits, combined with the provision of attentive, personalized care in a familiar setting, can greatly mitigate anxiety and confusion. This is even more crucial for those battling with cognitive challenges. Additionally, the advantage of having home care also prevents individuals from encountering other air-borne diseases in hospital environments which can be life threatening in older populations.

e. *Reduced Future Hospitalization Cost:* Finally, visits to home have demonstrated significant potential in reducing hospitalization costs. As noted in a study conducted by CareMore Health program in Connecticut, the approach of treating high-cost, high-needs patients, who typically account for a significant portion of healthcare spending, directly at their homes can be financially efficient. By investing in preventive measures and early interventions for these patients, CareMore achieved savings by averting expensive hospital stays. This home-based approach resulted in a decline in hospital admissions by 12.5% and emergency room visits by 27.2% over a 10-month period when compared to the previous year [18]. Such outcomes clearly indicate that attending to patients' health needs in their homes can result in better care outcomes while simultaneously reducing the financial burden on the patients.

f. *Cost of Implementation:* According to the statistics collected by Genworth, the cost of incorporating home care into the suggested customer journey is $27 per hour (per patient) [19]. This cost isn't very expensive considering a delay in consultation could lead to a significantly higher cost associated with exacerbating side effects.

II) **Virtual Care**

a. *Contextual Insights:* Virtual Care or tele-health offers a seamless interface between patients and healthcare professionals. Virtual care, through Humana's collaboration with the Center Well Specialty Pharmacy, provides a comprehensive and personalized interface between patients and healthcare professionals. By harnessing advanced technology, it not only ensures prompt consultations and reduced wait times but also facilitates immediate access to specialized pharmacists and nurses 24/7. These professionals work hand in hand with patients, providing crucial emotional support, financial assistance guidance, and expert

advice on medication management from CenterWell Specialty Pharmacy. This partnership could be particularly useful for clusters 2 and 3 having higher disability rates as it also provides telephonic reminders and doorstep delivery on pharmacy refills, eliminating the need for travel.

b. *Comfort from Home:* One of the pivotal advantages of virtual care is the comfort patients feel when consulting from their homes. This eliminates the need for potentially strenuous travel, reduces exposure to hospital-borne infections, and provides a sense of security. Patients can openly discuss their concerns without the intimidation of clinical settings, ensuring they feel heard and cared for.
c. *Priority Care:* Virtual care systems often have priority queues, ensuring that patients requiring urgent care are attended to promptly. This is particularly vital for those experiencing treatment side effects or needing a timely second opinion after a significant diagnosis.
d. *Utility in Detour-1:* Upon receiving a cancer diagnosis, time is of the essence. Virtual care ensures patients receive a second opinion promptly. With the emotional phase that follows such diagnoses, being able to connect with a healthcare professional without delay, and from the comfort of one's home, can be immensely reassuring.
e. *Utility in Detour-2:* When facing adverse drug events from treatments like Tagrisso, every moment counts. Virtual care ensures that patients receive immediate attention, guidance on symptom management, or any necessary adjustments to their treatment plan. This immediacy can be lifesaving and can dramatically improve the patient's quality of life during treatment.
f. *Reduced Future Hospitalization Cost:* Lastly, considering the growing emphasis on healthcare efficiency, telehealth presents substantial cost savings for both patients and providers. As highlighted by a 2017 Health Affairs study, telehealth visits can offer nearly a 50% reduction in consultation costs for certain ailments, such as acute respiratory infections [20]. Moreover, the analysis further underscores the affordability of telehealth for various conditions, with most consultations costing significantly less than traditional in-person visits.
g. **Cost of Implementation:** The cost of incorporating virtual care into the suggested customer journey is $65 per session (per patient) [20].

   The value of a virtual care session can range between $40 to $90 per session. We have assumed the average of the range to go forth with our analysis.

D. Recommendation – 3: Patient Support

"What is the problem?"

A huge part of beginning and surviving the cancer journey and coming out victorious is being able to trust and share your journey with others. Along with this, patients feel a lot better if they get up to date information about how to navigate this arduous journey through reliable sources.

- *Patients Have Questions:* Looking at the key topics of discussion from analysis in the previous section, we observed that the most relevant questions people ask are with respect to side effects, treatment options and effectiveness of Tagrisso. Not just this, we also see that individuals on the internet seek emotional support and reassurance from other people going through the same journey. This delineates a clear need for professionals that can answer these questions that patients are curious about and support them with issues of low morale.
- *Patient Sentiments Need to Be Heard:* As observed in the sentiment analysis section in our report, the most prevalent emotions shown by people on the internet revolve around feelings of 'Inquisitiveness' and 'Frustration'. What this tells us is that people use social platforms to enquire about various questions they have and express their sentiments on factors related to the treatment and/or the journey. Additionally, people are also enquiring about insurance related questions which is of utmost importance for Humana to address as it mitigates the chances of misinformation from unreliable sources.

"What is the solution and where do we intervene?"

I) Case Managers

a. **Contextual Insights:** The American Society of Clinical Oncology defines case managers as "educators and advocates for the person with cancer." Therefore, we advocate for assigning a personal case manager to every patient. This will make sure that all the patients have a first point of contact and can reach out to them about any issues that they might be facing. It is expected that all the other recommendations for patient support should go through the case manager as they are in the best position to help them.

b. **Use of Generative Artificial Inteligence (GenAI) in Personalized Patient Reports:** GenAI can empower Case Managers by analyzing historic data to generate personalized patient reports that seamlessly integrate financial insights, medical history, and timely alerts. Leveraging machine learning, it can identify patterns, ensuring proactive care. Additionally, GenAI's real-time alert system can notify Case Managers of any urgent issues through the Illness to Wellness journey, allowing for immediate intervention. As seen in Figure 23 below, the Case Manager can input any patient ID to get a summary of their personal information and pharmacy claims. Further, the case manager also has the liberty to ask follow-up questions to the chat-bot to obtain more nuanced information like "What medications is the patient taking apart from chemotherapy medicines and what side-effects have they been linked to?". To demonstrate the capabilities of the Patient History Assistant, we obtained the synthetic data available on Humana's Developer Portal and further embellished it by adding random drug class names, copay amounts, prices, etc. [21] More details about the technicalities are present in the appendix section.

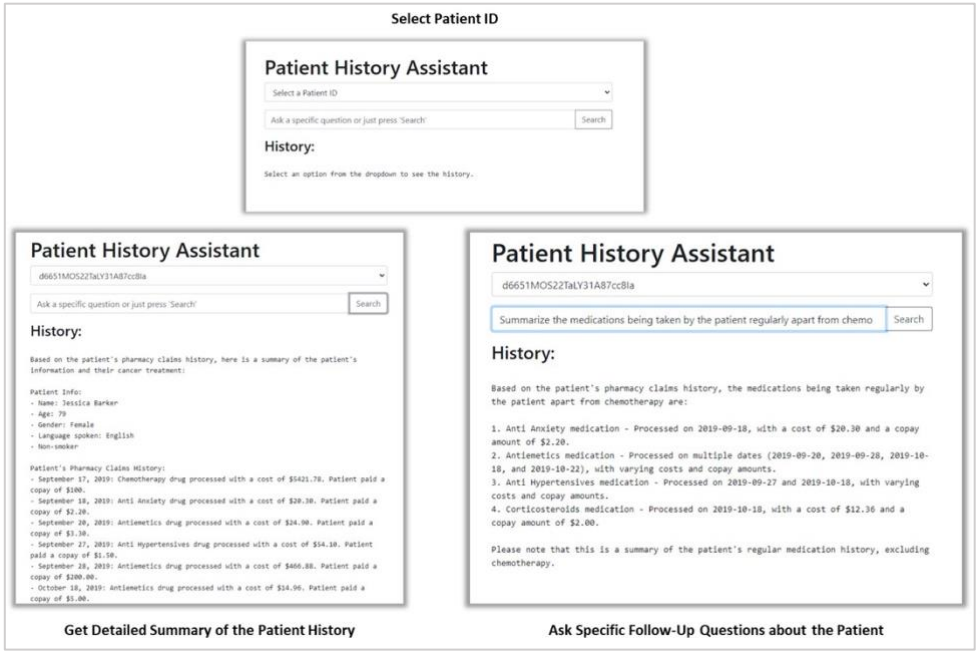


Figure 23:Patient History Assistant

c. *Utility at the beginning of the Journey:* The case managers help the patients understand the path of the cancer treatment and navigate the complex healthcare system. They can also give patients and their families crucial information regarding the diagnosis, available treatments, potential side effects, and what to anticipate throughout the course of treatment. They can use the Patient Assistant Bot to understand the patient's medical and financial history even before their first meeting.

d. *Utility in Detour -1:* The case managers should work with the patients in understanding their financial issues and medical concerns. At this point, they can refer the patients for a second consultation and appropriate support groups to ensure that the prospective patients sign up for the Tagrisso therapy.

e. *Utility in Detour – 2:* If a patient is hesitant to continue the treatment after the initial treatment, the case manager should first understand the issues of the patient and appropriately nudge them towards financial support options or Medical Re-consultation. The case manager can also provide psychosocial and emotional support. It is very important for the case manager to understand if the patient is facing a mental or emotional issue and appropriately direct them towards the corresponding support group. Lastly, Case Managers can use the Patient Assistant Bot to summarize medical history in case patients come back with signs of side effects. This could act as a first layer of information before a re-consultation or any change in financials is required.

f. *Cost of Implementation:* According to GoodRx, the cost of incorporating the role of a Case Manager into the suggested customer journey ranges between $90 to $250 per hour. We assumed a liberal estimate of $200 per hour (per patient) [22] to go forth with cost-benefit analysis. If the actual cost of case managers is less, it would lead to more savings overall.

II) Educating the Patient

a. *Contextual Insights:* In this case we suggest that educating the patient and making them aware of the effects of the cancer treatment should be one of the priorities for Humana. Johnson et al. (2021) [23] mentioned that in their sample size of 200 across different social media platforms 32.5% articles contained misinformation and 30.5% contained harmful information. This also falls in line with our topic analysis reiterated in the problems section of the patient support recommendation, where people are inquisitive about medical, financial and insurance related questions. This is where educating the patient and their caregivers becomes the utmost important so that they do not panic, try to self-medicate, and jump to inappropriate conclusions about their treatment.

b. *Utility at the beginning of the Journey:* Patients should understand the type and stage of cancer and the kind of diagnosis that they are signing up for. The patient should know the treatment options and the potential risks associated with each of these. This will help in the prevention of misinformation being spread.

c. *Utility in Detour - 1:* The case manager can work with the patient here to help them understand the treatment better. The patient should be educated about the nutrition and lifestyle changes, along with reinforcing the treatment timeline and providing the patient with the statistics of improvement through the Tagrisso treatment. The patient should be aware of the complete financial expectations so that they can plan accordingly in the future. To ensure that the patient's preferences and values are considered when choosing a course of treatment, shared decision-making should be promoted where the patient and the healthcare team actively collaborate.

d. *Cost of Implementation:* The cost of referring a patient to a comprehensive education program that covers everything from medical and financial questions, all the way to lifestyle changes would cost approximately $50 which is a one-time charge [24]. This cost considers mailing/print, educational material etc. Our recommendation is that patients are given proper lectures with guided information.

III) Support Groups

a. *Contextual Insights:* Support groups are essential for the well-being of cancer patients, allowing them to connect with peers facing similar challenges. Hoffman et al. (2021) [25] underscored their importance in enhancing the quality of life for lung cancer patients. The rise of online platforms presents opportunities for both patients and caregivers to find support without geographical constraints. It was observed from our topic analysis that the most common sentiment is ‘Hope’ which highlights the fact that Tagrisso treatment is largely successful and can be beneficial for a huge chunk of the patient population. However, as Walsh et al. (2021) [26] point out, these digital forums also pose risks of misinformation. It's

crucial to combine the accessibility of online resources with the credibility and depth of traditional support groups, ensuring patients and caregivers have trusted avenues for support and information.

b. *Utility in Detour - 1:* The patient and the primary caregiver is introduced to different support groups where they can learn and understand the progression of the Tagrisso treatment and feel empowered when they hear advocacy of the drug from the current users. Additionally, it will support the growth of peer mentoring, which offers encouragement and resilience by showing that people can overcome challenges and draw strength from their common experiences.

c. *Utility in Detour - 2:* Support groups offer reduction of isolation, and it also helps in talking to people who are also going through a similar phase. These groups are a safe space for communication where the patient can learn the success stories of different people, thereby feeling motivated enough to go through the treatment and not leaving.

d. **Cost of Implementation:** The cost of incorporating a Support Group into the suggested customer journey is $90 per hour (per patient) [27]. These costs include everything from staff personnel salaries, facility costs (electricity, rental cost etc.) to Administrative/operating costs (catering, office supplies, medical books etc.)

E. Cost Benefit Analysis:

I) Cluster-wise Detailed Cost Analysis: Figure 24 details out the cost analysis.

a. *Virtual Care and Home Care Hours Allocation by Cluster:* High-risk Seniors demand 80 hours of Home Care and 2 sessions of Virtual Vare due to their complex side effect conditions, ensuring continuous and in-depth support. The Economically Challenged group, also vulnerable to side effects, requires 2 sessions of Virtual Care and 60 hours of Home Care, striking a balance between their financial limitations and the heightened care needed due to potential side effects. Pre-dropout Patients and Persistent Patients, being less vulnerable to side effects in their therapy, are allocated 1 session of Virtual Care or less and 10 hours of Home Care, targeting essential interventions and routine assessments to sustain their treatment adherence.

b. *Support and Care Recommendations by Cluster:* Both High-risk Seniors and the Economically Challenged groups, given their intricate health challenges and vulnerability to side effects, are assigned 2 hours in support groups to foster community understanding and shared experiences. Pre-dropout Patients and Persistent Patients, with a more established treatment pathway, benefit from a concise 1-hour session, emphasizing crucial guidance and peer connections.

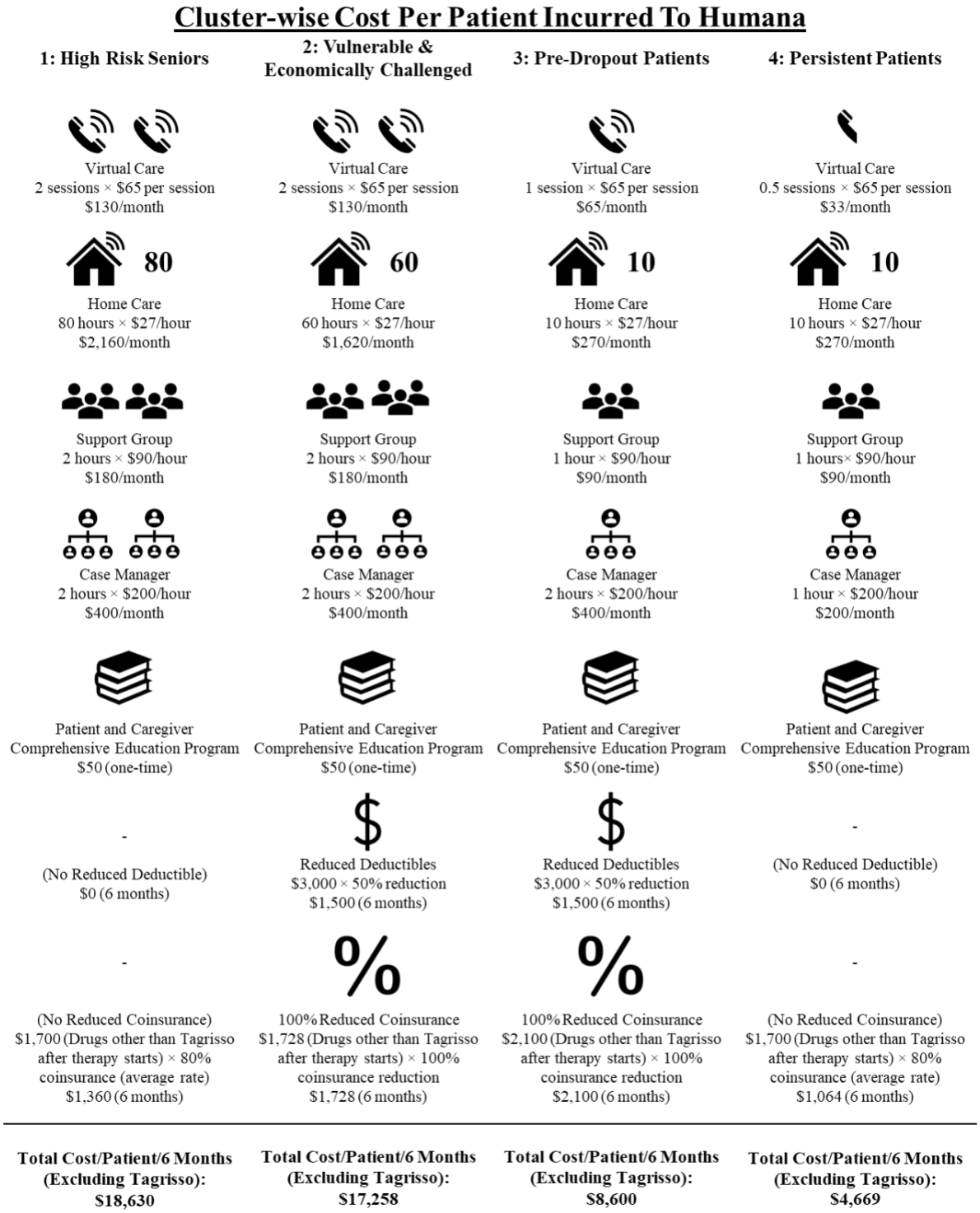


Figure 24:Cluster-wise per patient cost incurred

II) **Overall Cost-Benefit Calculation**

Our cost-benefit analysis method centers on a thorough examination of both retained and lost patients under our recommendations. The core of our revenue savings calculation is the comparison between two scenarios: first, where all Humana members diagnosed with NSCLC opt for alternative treatments like chemotherapy and surgery without the introduction of our recommendations and Tagrisso therapy; second, where 80% of these members continue with Tagrisso therapy due to our recommendations, while the remaining 20% opt alternative treatments.

a. *Step 1: Population and Patient Estimation*

   We began by determining the percentage of Humana Members relative to the US population. Using this ratio, along with US lung cancer statistics, we estimated the number of Humana members diagnosed with NSCLC.

b. *Step 2: Retention Rates*

   Our analysis assumed an 80% adherence rate, giving us both the projected number of patients who would stick with the Tagrisso therapy and those who might seek other options.

c. *Step 3: Cost Evaluation for Alternative Treatments*

   We delved into the costs associated with various treatment paths, like chemotherapy, radiation, and surgery. We calculated the average 6-month expense for a patient choosing alternative treatments, considering factors such as the proportion of patients selecting each treatment and inflation adjustments.

d. *Step 4: Recommended Solution Costs*

   We outlined the costs for each patient cluster, including the cost of Tagrisso. Using the estimated cluster sizes and the assumed retention figures, we gauged the potential expenditure for Humana should they follow our suggestions.

e. *Step 5: Revenue Savings Calculation*

   Lastly, we contrasted the cumulative costs tied to our advised solutions with the likely expenses if patients selected alternative treatments. This comparison yielded the potential revenue savings, underlining the financial merits of our recommendations.

Retained and Lost Patients Statistics for Recommendations		
US Population [28]	=	330,000,000
US Humana Members [29]	=	17,000,000
Humana Members to US Population Ratio	=	5.2%
Total New Cases of Lung Cancer in the US[30]	=	240,000
Proportion of Patients with NSCLC among Lung Cancer Cases [30]	=	80%
Total New Cases of NSCLC in the US: 240,000 × 80%	=	192,000
Number of Humana Members with NSCLC: 192,000 × 5.2%	=	9,891
Rounded Number of Humana Members with NSCLC	=	10,000
Assumed Recommendation Success Rate	=	80%
Retained (Continued Therapy) Customers: 10,000 × 80%	=	**8,000**
Lost (Discontinued Therapy) Customers: 10,000 × 20%	=	**2,000**

III) Cost of Alternative Treatment

Cost Components	$ Per Patient (per 6 Month)	% of Patients Taking Treatment (Early-Stage)	$ Per Patient Cost After Multiplying % of Patients
Chemotherapy [31]	$100,000	1% [36]	$1,000
Radiation [31]	$31,750	15% [36]	$4,763
Surgery One Time Cost [31] **(Assuming one surgery within first 6 months)**	$30,000	55% [36]	$16,500
Surgery + Chemo + Radiation	$161,750	16% [36]	$25,880
Chemo + Radiation	$131,750	6% [36]	$7,905
Other [32] **(No surgery, radiation, chemo)**	$59,000	7% [36]	$4,130
Additional Oncologist Appointments [33]	$2,032	100%	$2,032
Maintenance Medications & Drugs [34]	$24,000	100%	$24,000
Hospitalization Costs [35] **(Assuming 7.5 days in the first 6 months)**	$20,663	100%	$20,663

Notes: Some of the costs are adjusted through Inflation Rate Calculator[37]

Table 7: Cost of Alternative Treatment

Total 6 Months Alternative Treatment Cost (Per Patient) = $106,872 -------- (i)
Rounded Number of Humana Members with NSCLC = 10,000 ------- (ii)

Total Cost for Alternative Treatment = Total 6 Months Alternative Treatment Cost Per Patient (i) × Rounded Number of Humana Members with NSCLC (ii)

Total Cost for Alternative Treatment = $1,068,720,000

IV) Cost for Recommended Solutions

	High Risk Seniors	*Vulnerable & Economically Challenged*	*Pre-Dropout Patients*	*Persistent Patients*
Total Cost/patient/6 months (including Tagrisso - $90k)	$108,630	$107,258	$98,600	$94,669
Cluster Size	15%	10%	27%	48%
Patient Retained Number (80% - assumed above)	1,221	828	2,152	3,800
Estimated Total Cost for Each Cluster	$132,603,517	$88,765,241	$212,160,000	$359,742,200

Table 8: Cost for Recommended Solutions

Total 6-Monts Cost with Tagrisso i.e. Total Estimated Cost for All Clusters = $793,270,959
(80% of total customers assumed to be retained)

Total Cost for Missed Customer (Alternative Treatment) = 2,000 × $106,872 = $213,744,000
(20% of total customers assumed to be lost)

Total 6-Month Cost with Tagrisso for recommended solution = $1,007,014,959

Humana Revenue Savings

Revenue Savings = All Humana Members with NSCLC under Alternative Treatment - Total 6-Month Cost with Tagrisso for recommended solution.

Revenue Savings = ($1,068,720,000 - $1,007,014,959) = **$61,705,041**

10. Conclusion and Future Work:

To summarize, Humana stands to benefit from the integration of our predictive model. This model adeptly identifies individuals at risk of discontinuing therapy, aligning their situation with our Illness to Wellness framework. This allows for timely, tailored interventions based on each patient's unique state of discomfort. By prioritizing those at the highest risk, followed by individuals at moderate and low vulnerability, we can ensure that our intervention systems remain efficient and effective, without overwhelming available resources. Recognizing the deeply personal and often distressing nature of a cancer diagnosis, our approach remains unwaveringly patient-focused. Our recommendations emphasize the importance of understanding each patient's individual challenges and ensuring their concerns are genuinely acknowledged and addressed. Our comprehensive, three-tiered approach encompasses financial, medical, and patient support, ensuring that we holistically cater to every aspect of patient wellbeing.

Moving forward, we recommend that Humana incorporate regular feedback mechanisms during bi-weekly check-ins with patients. This valuable feedback can then be utilized to refine our Survival Analysis model, enabling us to pinpoint the exact moments when patients are most vulnerable and likely to discontinue therapy. Insights regarding their experiences with case managers, medical support teams, and financial plans can empower Humana to identify potential areas of concern early on. Finally, as an additional recommendation for the future, we suggest Humana launch a series of blogs and interviews showcasing patient success stories. Such narratives can act as beacons of hope for those at the early stages of diagnosis, painting a picture of resilience and recovery. By sharing these stories, Humana can foster a close-knit community of well-wishers, wherein patients and their families can draw strength from each other's journeys, further solidifying the sense of being part of a supportive and caring family.

11. Acknowledgement:

In closing, our heartfelt gratitude goes to Humana and the Mays Business School for providing us with the invaluable opportunity to delve into the compelling dataset related to cancer therapy discontinuation using *Tagrisso*. This endeavor not only illuminated the medical, financial, and operational challenges in the landscape but also enriched our understanding of how to meaningfully support patients. We deeply appreciate the chance to offer insights and solutions for issues that hold profound societal implications. This experience highlighted the significance of collaboration and the transformative power of data in bettering lives.

References:

[1] *Cancer facts & figures 2023*. American Cancer Society. (n.d.). https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/2023-cancer-facts-figures.html

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

[3] Padrnos, L., Bennett, R., Kosiorek, H., Dueck, A. C., Northfelt, D., Mikhael, J., Tibes, R., Khera, N., & Mesa, R. A. (2016). Living with cancer: An educational intervention in cancer patients can improve patient-reported knowledge deficit. *Journal of Cancer Education*, *33*(3), 653–659. https://doi.org/10.1007/s13187-016-1123-1

[4] SMITH, H. R. (2015). Depression in cancer patients: Pathogenesis, implications and treatment (review). *Oncology Letters*, *9*(4), 1509–1514. https://doi.org/10.3892/ol.2015.2944

[5] Free 2024 ICD-10-CM codes. (n.d.). https://www.icd10data.com/ICD10CM/Codes

[6] Abelson, R. (2023, July 19). *Insurers deny medical care for the poor at high rates, report says*. The New York Times. https://www.nytimes.com/2023/07/19/health/health-insurance-medicaid-denials.html

[7] Cass, A. (n.d.). *13 top reasons for claims denials*. Becker's Hospital Review. https://www.beckershospitalreview.com/finance/13-top-reasons-for-claims-denials.html

[8] Hall, P., Cox, B., Dickerson, S., Ravi Kannan, A., Kulkarni, R., & Schmidt, N. (2021). A united states fair lending perspective on machine learning. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.695301

[9] *What is adverse impact? and why measuring it matters*. hirevue.com. (n.d.). https://www.hirevue.com/blog/hiring/what-is-adverse-impact-and-why-measuring-it-matters

[10] Hummel, M., Edelmann, D., & Kopp-Schneider, A. (2017). Clustering of samples and variables with mixed-type data. *PLOS ONE*, *12*(11). https://doi.org/10.1371/journal.pone.0188274

[11] Le-Rademacher, J., & Wang, X. (2021). Time-to-event data: An overview and analysis considerations. *Journal of Thoracic Oncology*, *16*(7), 1067–1074. https://doi.org/10.1016/j.jtho.2021.04.004

[12] Dudley, W. N., Wickham, R., & Coombs, N. (2016). An Introduction to Survival Statistics: Kaplan-Meier Analysis. Journal of Advanced Practitioner in Oncology, 7(1), 91–100. https://doi.org/10.6004%2Fjadpro.2016.7.1.8

[13] Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of Medical Internet Research*, *15*(11). https://doi.org/10.2196/jmir.2721

[14] Prasad, A. (2022, June 17). *Council post: The role of patient sentiment analysis in the age of healthcare consumerism*. Forbes. https://www.forbes.com/sites/forbesagencycouncil/2022/06/16/the-role-of-patient-sentiment-analysis-in-the-age-of-healthcare-consumerism/?sh=27471f2a555f

[15] HealthPayerIntelligence. (2022, January 3). *Cancer patients in high deductible health plans face steep costs*. https://healthpayerintelligence.com/news/cancer-patients-in-high-deductible-health-plans-face-steep-costs

[16] *What is Coinsurance & How Does It Work?*. MetLife. (n.d.). https://www.metlife.com/stories/benefits/what-is-coinsurance/#:~:text=Insurance%20companies%20list%20percentages%20in

[17] T. Cornwell, "House Calls Are Reaching the Tipping Point — Now We Need the Workforce," *Journal of Patient-Centered Research and Reviews*, vol. 6, no. 3, pp. 188–191, Jul. 2019, doi: https://doi.org/10.17294/2330-0698.1719

[18] "How Home-Based Primary Care Can Reduce Expensive Hospitalizations," *Harvard Business Review*, May 16, 2019. https://hbr.org/2019/05/how-home-care-can-reduce-expensive-hospitalizations

[19] "Cost of Long Term Care by State | Cost of Care Report | Genworth," *www.genworth.com*. https://www.genworth.com/aging-and-you/finances/cost-of-care.html/

[20] "How Much Does a Telehealth Visit Cost?," *GoodRx*. https://www.goodrx.com/healthcare-access/telehealth/how-much-does-telehealth-cost

[21] [1] Humana Data Exchange (n.d.). Synthetic Data - Humana Developer Portal. Retrieved from https://developers.humana.com/syntheticdata

[22] "Medical Case Managers: How Do They Help Patients Save Money?," *GoodRx*. https://www.goodrx.com/healthcare-access/patient-advocacy/medical-case-managers-save-money

[23] Johnson, S. B., Parsons, M., Dorff, T., Moran, M. S., Ward, J. H., Cohen, S. A., Akerley, W., Bauman, J., Hubbard, J., Spratt, D. E., Bylund, C. L., Swire-Thompson, B., Onega, T., Scherer, L. D., Tward, J., & Fagerlin, A. (2021). Cancer misinformation and harmful information on Facebook and other social media: A brief report. *JNCI: Journal of the National Cancer Institute*, *114*(7), 1036–1039. https://doi.org/10.1093/jnci/djab141

[24] A. M. Stey, K. Fei, R. Franco, A. Mendelson, and N. A. Bickell, "Cost of comprehensive patient assistance program in early breast cancer patients," *SpringerPlus*, vol. 2, no. 1, Apr. 2013, doi: https://doi.org/10.1186/2193-1801-2-173

[25] Hofman, A., Zajdel, N., Klekowski, J., & Chabowski, M. (2021). Improving social support to increase QOL in lung cancer patients. *Cancer Management and Research*, *Volume 13*, 2319–2327. https://doi.org/10.2147/cmar.s278087

[26] Walsh, C. A., & Al Achkar, M. (2021). A qualitative study of online support communities for lung cancer survivors on targeted therapies. *Supportive Care in Cancer*, *29*(8), 4493–4500. https://doi.org/10.1007/s00520-021-05989-1

[27] A. M. Stey, K. Fei, R. Franco, A. Mendelson, and N. A. Bickell, "Cost of comprehensive patient assistance program in early breast cancer patients," *SpringerPlus*, vol. 2, no. 1, Apr. 2013, doi: https://doi.org/10.1186/2193-1801-2-173

[28] Wikipedia Contributors. (2019). *Demography of the United States*. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Demographics_of_the_United_States

[29] Rudden, J.. (2023) *Medical membership of Humana in the United States from 2008 to 2022.* Statista. Retrieved October 16, 2023, from https://www.statista.com/statistics/210651/total-medical-membership-of-humana/

[30] American Cancer Society. (2023). *Key Statistics for Lung Cancer. Retrieved October 16, 2023, from https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html*

[31] CostHelperHealth (n.d.). *Lung Cancer Treatment Cost*. Retrieved October 16, 2023, from https://health.costhelper.com/lung-cancer.html

[32] Boulanger, M., Mitchell, C., Zhong, J., & Hsu, M. (2022). Financial toxicity in lung cancer. *Frontiers in Oncology*, *12*. https://doi.org/10.3389/fonc.2022.1004102

[33] Dusetzina, S. B., Basch, E., & Keating, N. L. (2015). For Uninsured Cancer Patients, Outpatient Charges Can Be Costly, Putting Treatments Out Of Reach. Health Affairs (Project Hope), 34(4), 584–591. https://doi.org/10.1377/hlthaff.2014.0801

[34] Dorwart, L.. (2022). *How Much Does Lung Cancer Treatment Cost?* Verywellhealth. Retrieved October 16, 2023, from https://www.verywellhealth.com/lung-cancer-treatment-cost-5217734#:~:text=Here%20are%20some%20of%20the%20average%20costs%20for%20common%20lung,and%20marketplace%20insurance%20plans)13

[35] Holmquist, L., Russo, C. A., & Elixhauser, A. (2006). *Hospital Stays for Lung Cancer, 2006*. PubMed; Agency for Healthcare Research and Quality (US). https://pubmed.ncbi.nlm.nih.gov/21595130/

[36] Miller, K., Siegel, R., Jemal, A., & Dvm. (2016). *Estimated Numbers of Cancer Survivors by State as of January 1, 2016*. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/cancer-treatment-and-survivorship-facts-and-figures/cancer-treatment-and-survivorship-facts-and-figures-2016-2017.pdf

[37] CPI Inflation Calculator. (n.d.). *$20 in 2012 is worth $25.49 in 2022*. Retrieved October 16, 2023, from https://www.officialdata.org/2012-dollars-in-2022?amount=20#:~:text=The%20dollar%20had%20an%20average

Appendix:

The data for the below table has been taken with reference from Drugs.com. (https://www.drugs.com/drug-classes.html link last accessed on 10/15/2023 at 6:53 PM)

Type of Drugs	Side Effects Treatment
Analgesic Opioids	Used for pain relief
Corticosteroids	Often used to reduce inflammation
Dermatological Agents	Used for skin conditions like rashes
Antiemetics	Used to prevent nausea and vomiting
Antianxiety Agents	Help to reduce anxiety
Antipsychotics/Antimanic Agents	Used to manage psychiatric conditions
Tetracyclines	A type of antibiotic
Fluoroquinolones	Another antibiotic type
Hypnotics/Sedative/Sleep Disorder Agents	Used to manage sleep disorders and for calming effects
Anti-infective Agents	Help to fight infections

Table 1: Summary of Drugs along with Side Effect Treatment

Classification Report for the LightGBM Model *(Precision/Recall/F1 Scores at a threshold of 0.5):*

Target	Precision	Recall	F1 Score	Support
0	98%	95%	97%	1115
1	65%	83%	73%	117
Accuracy	94%			
Macro Avg Scores	81%	89%	85%	1232
Weighted Avg Scores	95%	94%	94%	1232

Table 2: Classification Report of the final chosen LightGBM Model

ROC Curve *(Light GBM Model):*

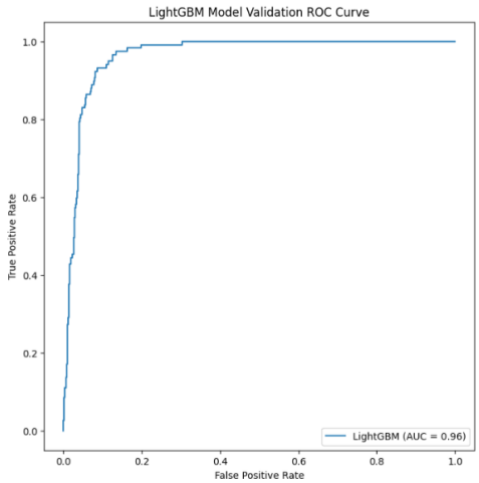


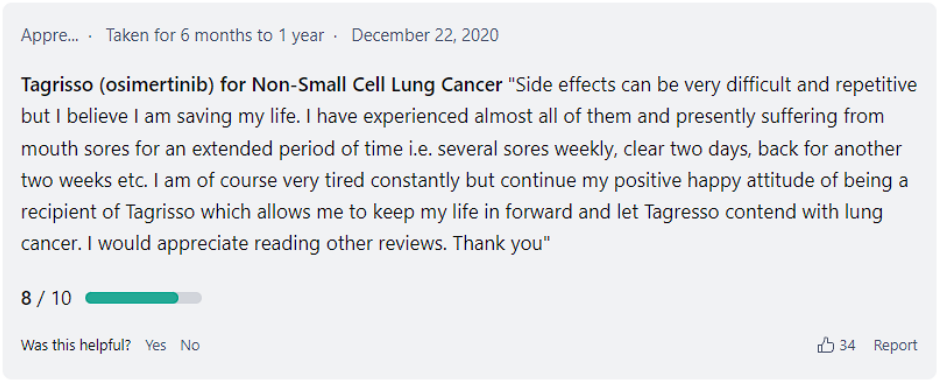
Figure 1: Receiver Operator Characteristic Curve for the LightGBM Model

Examples of Scraped Online Reviews:

https://www.drugs.com/comments/osimertinib/tagrisso.html

https://healthunlocked.com/lungcancer/posts/149659571/tagrisso-and-hair-thinning

https://reviews.webmd.com/drugs/drugreview-170434-tagrisso-oral



Generative AI-driven Patient Assistant Workflow

In order to demonstrate the capabilities of the Patient History Assistant/Bot, we obtained the synthetic data available on Humana's Developer Portal (https://developers.humana.com/syntheticdata) and further embellished it by adding random drug class names, co-insurance amounts, prices, etc. Post this, we utilized Open AI GPT3.5 Model with Prompt Engineering and Retrieval Augmented Generation techniques to create an app that could find and respond to user queries about the patient. The sample app was built using Flask. The implemented flow is supposed to take two inputs from the user – the unique patient identifier for fetching patient information and pharmacy claims information based on the id, and an additional query that allows the case manager to ask specific follow up questions to the bot about the patient. This input combined would be used to query the relevant patient information which would be summarized using the GPT3.5 Model and presented to the case manager.