PREDICTING LONG TERM OPIOID THERAPY (LTOT) OUTCOMES FROM INSURANCE CLAIMS DATA USING A RANDOM FOREST APPROACH

Humana Case Competition
October 2019

Long term opioid therapy (LTOT) is becoming increasingly common in the United States, at a high risk to patients and at a high cost to the medical system — including insurance providers like Humana. As such, accurately predicting whether or not someone is likely to become an LTOT patient at the time of an opioid naive prescription opens many opportunities for insurers, doctors, and patients take steps to better understand — and avoid — LTOT outcomes. Using four years of longitudinal insurance claims data from 14,000 Humana members, we built a random forest model that can predict at the time of opioid naive prescription whether or not a member will become an LTOT patient with an AUC of 0.916. In addition to prediction, our model offers interpretable insights that allow us to formulate recommendations related to sharing information with prescribers, Humana-lead opioid policies, and data tracking practices.

1 INTRODUCTION

Over the past two decades, the prevalence of patients on long term opioid therapy (LTOT) has increased considerably. Although LTOT originated as a way to manage the pain of patients with chronic conditions, the rapid increase in LTOT patients — despite a much smaller increase in the prevalence of pain conditions — suggests that there has been a fundamental change in prescriber behavior regarding pain management in the United States.

Prescribing opioids to a patient in a long-term manner is a non-trivial decision. LTOT patients have an increased risk of opioid overdose, abuse, and misuse, as well as other adverse health outcomes. Given these significant risks, it makes sense that many members of the healthcare system — such as physicians, pharmacists, and insurers — would like to identify the drivers that lead a patient to LTOT, as well as a patient's likelihood to require LTOT after their first opioid naive prescription. Doing so could allow for more informed decision making at the time of prescription, as well as earlier intervention to prevent a patient from becoming a long-term user of opioids.

At the heart of this medical/business need is a prediction problem. More specifically, the core predictive question is as follows: given a patient's documented medical history, what is the probability of them becoming a user of LTOT after today's naive opioid prescription? Given the rich longitudinal data provided by Humana that documents several years of medical history for thousands of patients, we built a predictive model to answer this very question. Our model was trained on members in the dataset for whom we know their LTOT outcome, and predicts the status of a new patient by considering their past claims. Our subsequent analysis documents our data preparation, model development, model findings, and business recommendations.

2 DATA PREPARATION

2.1 Creating Outcome Labels

A critical initial step of our data preparation process was creating the outcome labels for each patient and their opioid naive events. While this information was captured implicitly in the data, we had to extract these binary outcomes to create explicit labels to train our classification model. This label creation process had two main components: identifying opioid naive events, and identifying whether or not an opioid naive event resulted in LTOT.

In order to identify all opioid naive events for each patient, we subset the initial data to only look at "RX Claim - Paid" events. Among those events, we identified which prescriptions were opioids by looking at the drug class; all drugs that fell into the following five categories were marked as opioids — opioid agonists, opioid combinations, opioid partial agonists, opioid antagonists, antiperistaltic agents. We then checked each opioid prescription for each patient against the opioid naive criteria. If the prescription day was equal to zero, or if the difference between the current prescription day and last prescription day was greater than 90, an opioid prescription was denoted opioid naive.

Once we identified opioid naive prescriptions for each patient, we had to check if these opioid naive prescriptions resulted in LTOT. To do this, we needed to check if the patient had an opioid on hand 90% of days within the 180 days following the opioid naive prescription. One challenge that had to be overcome in checking this condition was that several patients in the data receive multiple opioid prescriptions on the day of their opioid naive events. To overcome this challenge, we introduced a new variable, "opioid_until", which checks the patients' opioid availability for the future and avoids double counting the multiple opioid prescriptions in labeling.

The final output here is a list of all opioid naive events for each patient, the day of the event, and whether or not it resulted in LTOT. After creating our final labels, we checked the percentage of LTOT outcomes in our data against the 30 percent prevalence rate that Humana said we should observe. Our labels had a 29.7 percent prevalence of LTOT, suggesting our labels are correct. Additionally, our average number of opioid naive events per person matched the 1.6 that Humana had observed. The final LTOT labels are used as the dependent variable in our classification model, which will be discussed later in this analysis.

2.2 Data Exploration

We began our analysis by doing detailed data exploration to better understand the nature of the Humana claims data and the data irregularities. Below, we will highlight our findings that are most relevant to this analysis.

Member Missing an Opioid Naive Event:
We identified one member in the data who did not have an opioid naive qualifying event on or after Day 0. This individual can be identified by their member ID, "ID14698922966". Based on the definition of the data provided by Humana and the goal of the prediction problem, all members of the dataset should have had an opioid naive event. As such, we decided to consider this person an anomaly, and exclude them and their data from our model training dataset.

NA's in New Provider Event Type:
In all the data provided, the only event type for which all columns were NA for all observations was the "New Provider" event. This data quality issue prevented us from being able to better understand the details of New Provider events, such as the associated Diagnosis and Place of Treatment. However, we proceeded with an assumption that despite the lack of details, an instance of this event still accurately indicated a member had a new provider. We used the count of these instances, rather than their details, in our feature generation process.

Opioid Naive Events Before Day 0:
The goal of our model was to predict LTOT outcomes for data on or after Day 0. However, in our data exploration we discovered that many patients have rich histories of opioid use (including LTOT instances) before their Day 0 opioid naive event. This finding had two main implications. The first implication was that we had to adapt the way we built our outcome labels so that opioid naive/LTOT

events before Day 0 were disregarded. The second implication was that we could use historical opioid events to build relevant, opioid-specific features for our model.

2.3 Feature Engineering

The patient data provided by Humana was longitudinal, meaning that for each patient, we had many observations over time. However, to build a binary classification model using our desired approach, we needed to have the data structured as one observation per patient, per opioid naive event. Most patients have multiple opioid naive events (the mean number of opioid naive events per patient is 1.67, and the max is 7), suggesting that each member will have multiple observations in our final training dataset. Getting the data into this format required extensive data manipulation and feature engineering.

Each of our features is built off of all of the data from the beginning of the patient's history, up to the date of the opioid naive event that we are trying to predict the outcome for. For example, for a patient's opioid naive event on day 0, all of their longitudinal observations up to and including day 0 will be included in the calculation for each feature. However for an opioid naive event on day 150, all longitudinal observations up to and including day 150 will be included in the feature calculations. The reason for this setup is it models the prediction mechanism for our validation data and real-life scenarios.

We built features in five main categories. Below, we will walk through these main categories and highlight a sample of features within each. For the full list of features built, please refer to our appendix in Section 7.

<u>Prescriptions:</u>
A large number of the features we created were related to the prescription events of members. We decided to create detailed features related to prescriptions because a member's prescription history contains rich information about his/her general health status, interactions with medical providers, and opioid history. Four subcategories of prescription features emerge within this domain.

The first subcategory is features related to opioid prescriptions. To build features related to opioid prescriptions, we subset the data to only look at "RX Claim - Paid" events for an opioid. From this subset of data, we created a wide range of historical features related to the number of past opioid prescriptions a member has received, the past average morphine milligram equivalent (MME) the member receives, the type of opioids the member has previously been prescribed, and the reasons for their previous opioid prescriptions. We also built these features over different time horizons in an attempt to capture the time series information in our data. Additionally, we created several features related to the opioid naive prescription event, including the MME of the qualifying prescription, supply count of the qualifying prescription, and the way in which the qualifying prescription was administered.

The second subcategory is features related to non-opioid prescriptions. To build these features, we subset the data to only look at "RX Claim - Paid" events, but this time for only non-opioid prescriptions. Some of the features we built in this category include the total number of other prescriptions for a

member, whether or not they were on other pain medications or benzodiazepines, and whether or not they were on medications for anxiety, depression, or other psychological conditions. We derived inspiration for many of these non-opioid prescription features from outside literature, which suggested that other pain medications, benzodiazepines, and mental illness were associated with higher LTOT risk (Sullivan 2006; Lautieri 2019).

The third subcategory is features related to rejected prescription claims. These observations come from the "RX Claim - Rejected" event category. We created features for a member's total number of rejected claims as well as the number of rejected claims that fell into each of the different status categories. We intended for these features to capture any correlation between rejected claims and LTOT outcomes.

The final subcategory is features related to mail prescriptions. Using the "RX Claim - New Mail Order" event type, we created features to capture the number of mail prescriptions over different time horizons in attempt to capture any relevant correlation between individuals receiving prescriptions via mail and LTOT outcomes.

Diagnosis History:

Another major category of features we created were diagnosis related features. These features came from the following event types: "New Diagnosis - CAD", "New Diagnosis - Diabetes", "New Diagnosis - Hypertension", "New Diagnosis - CPD", "New Diagnosis - CHF" and the "Fully Paid Claim" events. Each of these categories capture rich information about the medical status of the members of the dataset, and as such, we felt they would be helpful predictors of LTOT. For each member, we created variables based off of the number of events they had in the CAD, Diabetes, Hypertension, CPD and CHF data. We also built these features over several time horizons to try and capture signal about the sequencing of diagnoses. Although opioid literature doesn't specifically suggest that these five conditions have been shown to be related to LTOT outcomes, they constituted a large portion of the data provided by Humana and have the potential to lead to pain disorder, so we decided to include them as features in our model.

For the "Fully Paid Claim" events, we built specific features based on outside literature. Attribute 1 of these events contained detailed diagnoses as strings of text. By searching for keywords using Python string processing abilities, we were able to extract whether or not a member had diagnoses that come up frequently in the literature as related to LTOT outcomes. We searched for diagnoses related to bipolar disorder, anxiety, psychosis, suicide, self-harm, pain, depression, tobacco abuse, alcohol abuse, cocaine abuse, amphetamine abuse, opioid abuse, cancer, and schizophrenia and created features based off the frequency that these words appeared in the diagnoses of members (Sullivan 2006).

New Provider:

As mentioned in Section 2.2 (Data Exploration), the "New Provider" events contained all NA column values, so the only features we built based off of this event were the counts of events per member over different time horizons.

Surgery:

The final category of features we created are related to "Surgery" events. Opioids are often prescribed for post-surgery recovery and therefore this segment has high predictive potential. We built features to capture the number of surgeries a member had over different time horizons, as well as features to capture the type of facility where their procedures took place.

In aggregate, we created 118 features to summarize the longitudinal data for each opioid naive event observed in the data after and including Day 0.

3 PREDICTIVE MODELING APPROACH

3.1 Model Requirements

After building features, creating outcome labels and transforming our final dataset into one row per opioid naive event, we were ready to build a binary classification model to predict whether or not an opioid naive event would result in an LTOT outcome.

To begin, we split our dataset into two parts, a training dataset which contained 80 percent of the observations, and a test set which contained 20 percent of the observations. We performed this split randomly, but stratified the data to ensure that the same proportion of LTOT = 1 observations were present in the train and test data. The training data is what we used to build our model, and the test data is what we used to compare models and validate our final model's performance.

We began by evaluating our training data in order to get a sense of the right choice of binary classification model. Our training data has several characteristics that impact this decision:

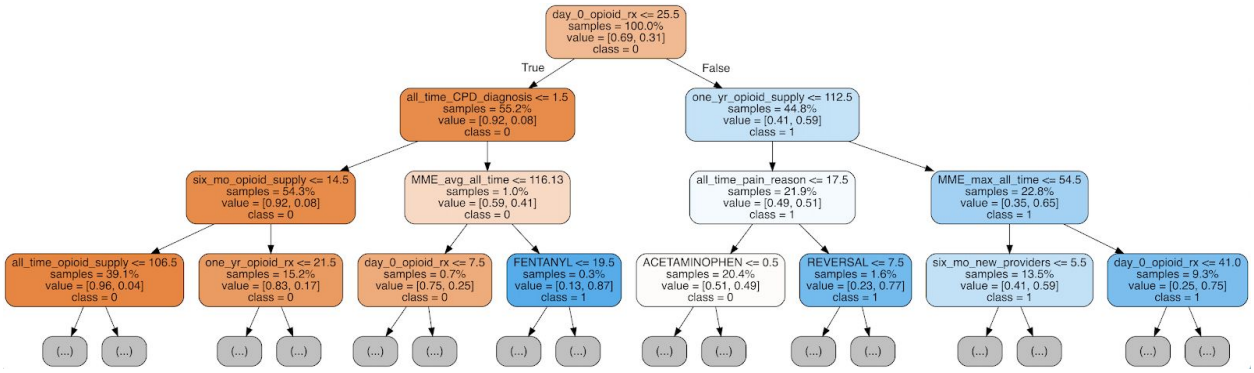
- The dataset is "wide", exhibiting high dimensionality due to the large number of features we created. This high dimensionality also implies that many individuals are likely to be outliers on at least some dimensions.
- Many of our features are highly correlated. This is particularly true for the features we built over different time horizons. For example, we have a feature that captures the number of opioid prescriptions for a member in the past 6 months, and another variable that captures the number of opioid prescriptions for a member in the past 1 year. These tend to be highly correlated.
- Our final training dataset has 30 percent of observations with an LTOT = 1 outcome label, and 70 percent with an LTOT = 0 outcome label. This difference in percentages suggests we have a moderate class imbalance issue in our data.
- The features in our data had many different ranges and scales. For example, our variable that captures the number of times a member has been LTOT before has a very narrow range (all numbers between zero and five, whereas our variable that captures the max MME a member has ever recieved can range from ten to thousands.

3.2 Final Model

Taking all of the factors above into account, we decided that Random forest was the most appropriate model for our data. Random forests are an ensemble method for classification, built using many individual decision trees. Random forests use the output of each individual tree to "vote" on a single final outcome for each observation in the data.

The architecture of random forests make them well suited to handle all of the characteristics of our data referenced in the previous sections. Firstly, random forests handle dimensionality and highly correlated variables well because each individual decision tree is allowed to split only on a random subset of features. In essence, the dimensionality problem is not as relevant for random forests than other models because each tree is trained on a dataset of lower dimensionality, but in aggregate they capture the full feature set. The correlation problem is less relevant than other models because each tree is trained on a *random* subset features. Second, because random forest is a tree-based model, it is less sensitive to class imbalanced data that other classifiers like logistic regression. Third, random forests require little preprocessing, and can intake data with different scales. Fourth, random forests handle outliers well because each individual CART tree is built off of a bootstrapped sample of data, reducing the effect of outliers and enhancing the models "wisdom of the crowds" abilities.

In addition to being capable of addressing our data's challenges, a random forest model proved to have many other advantages for this problem. Relative to the accuracy level achieved, random forest models have short training time. Random forests are also computationally efficient, and would scale well to all of Humana's members if our model would be applied to a large sample of Humana's data. Additionally, tree-based models are also typically easier to explain to stakeholders, particularly those lacking a strong data science or statistical background.



The image above is an example of the top of a single decision tree from our random forest model. We have included it to show the way in which the decision process of a random forest can be easily visualized and interpreted. Colors indicate the majority of classes in that node, orange being patients without LTOT and blue being LTOT patients. We can visualize the percentage of LTOT patients in each split as well, in the value details in each node.

Lastly, random forests provide a meaningful interpretation of important variables though the Gini Impurity Measure. We will explain this variable importance measure (and the interpretation of the variable importances in our model) in Section 4.1.

3.3 Exploring Other Models

Before making our final decision that a random forest was the most appropriate model, we investigated other classification methods that could potentially help us to interpret the results easier, or achieve higher accuracy in detecting future long term opioid users. We started with a simpler classification model, logistic regression, to have a baseline algorithm for our random forest model. After building a logistic regression model, we realized that it was not robust and accurate due to the high dimensionality and high correlation between the variables in the dataset.

Support vector machines offered another approach for this classification task. However, the training time increased significantly and the accuracy was far worse before hyperparameter tuning. Considering the time tuning would have taken, we decided not to move on with support vector machines in favor of focusing on a more interpretable and computationally efficient model.

We also tried boosting models. We fit adaptive boosting and gradient boosting models to our training data and compared our evaluation metrics with the results of the random forest. Since boosting is a tree-based algorithm, we achieved accuracy, AUC and F1 scores similar to those in our random forest model. We tuned learning rate, depth of the trees and regularization parameters to further improve the accuracy. Nonetheless, the marginal improvement in the evaluation metrics were nearly negligible. We determined that the very small improvement in performance was not worth the disadvantages of increased model complexity.

The type of final model we explored was a neural network. We tried different neural network architectures by changing the depth of the layers, the type of activation functions and the hyperparameters. Similarly to the case with boosting, we decided the interpretability of our random forest was preferable to the nearly negligible improvements in performance metrics in this model.

After all of these attempts, we were confident that a random forest was the right model for this classification problem.

3.4 Generalizing Approach for New Data

In order to predict the probability of a patient becoming LTOT in the holdout dataset for our final submission, we had to replicate our feature engineering work and run the random forest model on the holdout data. Because the naming convention for columns was different in holdout dataset we had to adjust the column names in the holdout set to avoid any errors. Additionally, in the holdout set, we were only attempting to predict one Day = 0 opioid naive event per person, meaning we did not need to transform one member into several observations based on their number of opioid naive events. Other

than these small differences, our feature engineering code easily transferred over to the holdout set. It was then simple to run the random forest model on this new data, and generate our CSV of predictions for submission.

4 MODEL EVALUATION

4.1 Model Metrics

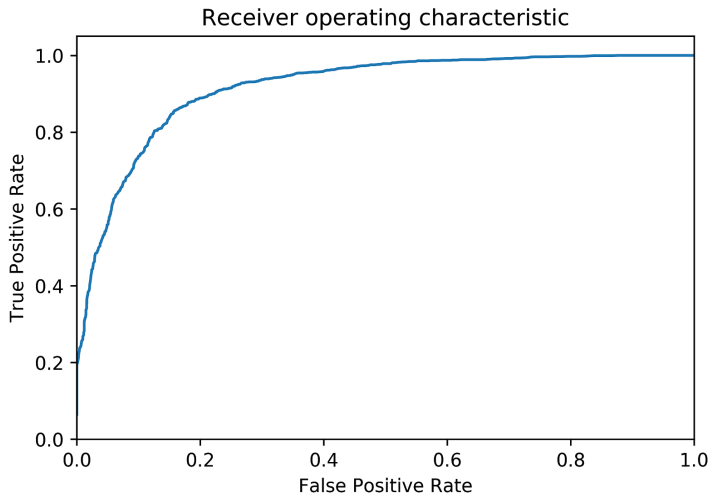
The main metric of interest for this case competition — as defined by Humana — is the area under the curve (AUC). In this section, we will discuss our model's performance on our test set in terms of AUC as well as other evaluation metrics that we felt were important to assessing the quality of our model.

Establishing Baseline Model Performance:

Before evaluating the performance of any modeling approach, it is critical to establish a baseline model for comparison. Baseline models are often naive modeling approaches that establish a lower bound for model performance. In this prediction problem, a baseline model could be to simply predict the outcome that is most prevalent in the training data. Because 70 percent of training observations had LTOT = 0 and 30 percent had LTOT = 1, LTOT = 0 is the most prevalent class in the training data. If we predicted all observations in our test set to be the most common class (LTOT = 0), our accuracy would be 70%. This accuracy represents the baseline accuracy which we will compare our model against.

AUC:

The area under the curve (AUC) is a measure of the area under the Receiver Operating Curve (ROC). The ROC curve is a plot of a model's true positive rate relative to its false positive rate, highlighting the model's ability to distinguish between classes. AUC ranges from 0.5 to 1.0, with values closer to 1.0 corresponding to more predictive classification power.



The curve above represents the ROC of our model on the test data. As we can see, the curve is close to the upper left corner, which indicates a high predictive power. It is significant to visualize the ROC for understanding the trade-off between the true positive rate and false positive rate.

Our final model had a .916 AUC on our test data. This strong AUC suggests that our model does a sound job of differentiating LTOT = 0 observations from LTOT = 1 observations. It is important to note that AUC also has a probabilistic interpretation. An AUC of .916 implies that given the predicted score of two members — one whose outcome is truly LTOT = 1, and the other whose outcome is truly LTOT = 0 — our model will correctly assign a higher score to the LTOT = 1 member 91.6% of the time.

Confusion Matrix and Accuracy:

A classification model's accuracy is defined as the number of observations who are predicted to be in their true class divided by the total number of observations. Our random forest had an accuracy of 84.8% on our test data. It is important to compare this accuracy to the baseline accuracy of 70% that we established earlier in this section. Our model performs 14.8% better than the naive model that predicts all unseen data to be part of the majority class (LTOT = 0), and 34.8% better than random chance.

The table below is the confusion matrix for our test data:

	Predicted LTOT = 0	**Predicted LTOT = 1**
Actual LTOT = 0	3221	375
Actual LTOT = 1	432	1205

An interesting thing to note in our data is the relative balance between false positives and false negatives. False positives are observations we predicted to be positive but are actually negative. False negatives are observations that we predicted to be negative but are actually positive. In our training data we had 375 false positives and 432 false negatives, suggesting that our model has found a decent middle ground in making tradeoffs between false positives and false negatives. In this use case, it is important to note that false negatives are likely more costly that false positives. This is because it is far worse to miss a patient who will become LTOT — and as a result fail to deliver them resources and interventions — than it is to provide resources and interventions to a patient who may not have needed them.

Precision, Recall, and F1 Score:

Overall accuracy summarizes our overall performance, but fails to identify more specific areas in which our model is stronger/weaker. Precision, Recall, and F1 Scores at the aggregate and class levels are helpful tools for understanding how our model performs in more specific circumstances.

	Precision	Recall	F1 Score
LTOT = 0	0.88	0.90	0.89
LTOT = 1	0.77	0.74	0.75
Average (Total)	0.85	0.85	0.85

Precision is the fraction of observations we predicted to be a certain class that actually belong to that class. Here, our precision is consistently strong across both classes, with stronger performance on class 0 than 1. Recall is the fraction of observations that are actually a certain class that we predicted to be in that class. Once again, we see a better recall score for class 0 than class 1. This better performance on class 0 is to be expected, since a larger proportion of our training data belonged to this class.

The F1 Score is the harmonic mean between precision and recall. F1 score can be interpreted as a summary measure of how precise and robust your classifier is. The F1 score was stronger for class 0 than class 1 (for the same reasons as mentioned above), but on average was 0.85, indicating a robust model.

4.2 Variable Importances

In order to provide useful insights and recommendations to Humana based on our analysis we utilized the variable importance feature made available for random forest models in Python Scikit-Learn. The variable importance function implements a Gini impurity measurement — when aggregated, a measure of how much information each feature in the model provides in classifying a new observation — to determine which features are the strongest predictors in the model.

Our model uses information up through the opioid naive event prescription(s) to predict whether a member will be administered LTOT. Our features can therefore be interpreted in three main groups: those features that represent a member's past opioid-related medical history, those features that represent a member's past non-opioid-related medical history, and those that pertain to their opioid naive prescription(s). Our most important features pertaining to the first category are some of the most predictive: a member's documented history of opioid prescriptions all-time and in the 6 months, 1 year, and 2 years leading up to the opioid naive event, the average and maximum MME of all opioid prescriptions leading up to the opioid naive event, the number of opioid prescriptions with reason code "pain", and whether or not a member has been LTOT before. It is largely intuitive why a member's previous opioid history would be relevant in predicting whether they would eventually require LTOT. Someone who has had LTOT before or who has a history of opioid use is more likely to have LTOT in the future because they already have patterned opioid behavior. We can also posit a positive correlation between high MME dosage and severity of pain or illness which may necessitate long term opioid use, and it has also been shown that higher-MME prescriptions carry higher risk for addiction and opioid seeking (Phillips 2017). The frequency of "pain" as a prescription reason is less transparent as a predictor of LTOT, but could imply that because pain is being treated as opposed to another condition, the pain could be chronic with an unclear cause. According to our research general pain is a known symptom

described by individuals actively seeking an opioid prescription, and this could contribute to this feature's importance (though this is certainly not an overarching statement about all members with chronic pain symptoms) (Kaye et al. 2017).

In addition to opioid-related historical predictors, there are several features that are not necessarily opioid related that also have high importance in our model. These include the total amount of claims both opioid and non-opioid related recorded for the member, the total amount of historical non-opioid prescriptions recorded, the total number of rejected prescriptions regardless of the prescription reason code, and the total number of providers documented for the member. The number of non-opioid prescriptions with reason code "pain", in particular NSAIDs prescription count, and the number of prescriptions for anxiety, depression, and other psychological disorders, in particular the Benzodiazepines prescription count, are all highly relevant predictors of future LTOT. There is potentially a negative correlation here between the number of total claims and total prescriptions a member has and their health status. Someone with an extended history of health issues is more likely to have a variety of health interventions in the future, one of which may be LTOT. Count of rejected prescriptions is harder to interpret given the limited insight provided by the rejection reason codes, but one plausible interpretation is that a patient that actively seeks medications that cannot be fulfilled/refilled for various reasons may be more likely to attempt to extend their opioid prescription (with variable success). A high number of prescribers can indicate a sick person with comorbidities, but based on our research can also point to someone who is actively seeking opioids (Mayo Clinic Staff n.d.). A history of pain and strong non-opioid pain medication like NSAIDs likely has a similar interpretation to opioid pain-related history in that it may imply a continued need for/interest in pain treatment. Finally, our research indicates that psychological disorders like anxiety and depression as well as other psychological disorders are predictors of opioid use, and specifically Benzodiazepines — a medication used to treat anxiety — are highly addictive and are often abused in tandem with opioids (Lautieri 2019).

Our final category — events on the day of an opioid naive prescription — also has several features that are of importance in predicting LTOT. Our most important feature in predicting LTOT falls into this latter category: the prescription coverage duration in days for the qualifying prescription. MME on the qualifying event, the number of opioid prescriptions on the qualifying event, and the unique number of types of opioids on the qualifying event (and in particular one being from the agonist subgroup) were also predictive of LTOT. The effect of number of days of initial opioid prescription on long-term opioid use is well-documented and it is therefore not surprising that this feature is of such high importance (Phillips 2017). The explanation for MME, number of opioid prescriptions, and unique types of opioids all have a similar interpretation to that provided above for historical data — in essence, these features indicate prescription strength and complexity. Similarly, opioid agonists fully activate the opioid receptors in the brain, and have the most potent impact of all the opioid subgroups (Trescot 2008). The strength and duration of an opioid prescription can positively correlate with both the severity of the members' illness, and also their increased likelihood for addiction.

5 RECOMMENDATIONS

Humana has a unique and comprehensive perspective on member interaction with the medical system via their claims data. Humana has the potential not only to provide information to medical professionals and researchers to help them make/develop informed opioid prescribing regimes, but also to enact policy to shape the prevalence of long-term opioid use in the United States. We have structured our recommendations built from our analysis into three main strategic areas: recommendations pertaining to information sharing, recommendations for Humana-lead opioid policy, and a set of more general recommendations for Humana in terms of their data tracking practices. Applying a holistic perspective of both prevention and treatment of opioid dependence is essential to reduce overall LTOT incidence.

In the variable importance section we described a series of opioid- and non-opioid-related features that were significant in predicting LTOT. Within the confines of medical data privacy laws, Humana should share as much information as possible with providers about members' medical claims history so that they can make informed decisions when making prescribing decisions. This is particularly important for the features that have high importance in our model — things like past opioid history, past pain treatment, and diagnosed psychological and substance abuse disorders. While providing data regarding these specific areas could be invaluable to a prescriber in their decision-making process, providers may have difficulty comprehending the extensive panel data that exists for most members. Humana should therefore also consider providing an LTOT risk score for each patient based on a predictive model similar to what we have built for this case competition. This would ideally go beyond the safety flags that the CMS Opioid Task Force already mandate to provide more comprehensive information to providers, even when a patient's history doesn't necessarily meet the requirements for a formal flag (CMS 2017). It is important that the risk metric provided isn't modeled via a black box method — prescribers must be able to understand the significant contributing factors to a member's particular risk score updated with each claim log so that it can be used in tandem with their medical expertise to make treatment decisions. If a tree-based model is employed, outputting a report of the top variable importances from the model will key the prescriber in to areas of the member's medical history that make them at risk for LTOT. Humana has an important role in educating providers about the risks of opioids and the alternatives that exist for pain mediation. In addition to sharing information with providers, Humana is well-positioned to make an impact on the opioid crisis by sharing data via the Predictive Learning Analytics Tracking Outcome (PLATO), a governmental initiative to pool insurance provider data. The business impact of cleaning and disseminating claims information and refining a predictive model for members is relatively low and shouldn't require Humana to make any major long-term adjustments to their operational practices. Humana has rich medical claims data that when used appropriately can provide incredible insight into the proportional increase in LTOT and novel ways to curb dependence.

Information sharing and opioid incident flagging for relevant parties is extremely important, but Humana is also in a position to develop opioid-related policy to more formally guide prescribers in their network and to transform their own policy to be a leader in bringing cultural change in pain management in the insurance industry. Based on our model and the extensive research being done in the field of long term opioid use, the decisions made by prescribers on the naive event date can have a profound impact on members' long term relationship with opioids. As our model identified, the number of days covered by

the prescription on the naive event date, the daily MME, and the type of opioid are all predictive of LTOT. These are all trackable prescription points that Humana can build policy around. Humana can mandate a maximum prescription day coverage count based on the opioid type and MME, a maximum daily MME based on health condition, and a ceiling for the number of opioid prescriptions a patient can get within a particular time period. Humana can also enforce checks with providers in addition to prior authorization for opioid prescriptions, instating a system to make sure they are monitoring members with opioid prescriptions and ensuring that they aren't seeing multiple providers for the same medical issues or filling prescriptions at multiple pharmacies (including mailed prescriptions, the count of which were moderately predictive of LTOT in our model). Developing tracking, warning, and reporting mechanisms internally and with third-party auditors for providers that prescribe opioids outside the realm of recommended practices and members that are high-risk for opioid abuse are a low-cost way to stay on top of opioid use growth.

Humana can also do its part in the insurance policy arena to support members through coverage of preventative measures, non-opioid pain management solutions, and opioid addiction counselling. Prevention and treatment must be supported in tandem to contain and improve upon the LTOT incidence rates, and while this section of recommendations is the most significant in terms of business impact for Humana, it also has great potential for long-term cost savings. Based on our model findings and ongoing opioid research, a great deal of pain has psychological origins, and dealing with that root cause can have a profound impact on the need for pain treatment. Broad-based access to psychological treatment can both prevent causes of pain (encouraging healthy lifestyles and prevent comorbidity) and help opioid-dependent members deal with their addictions (Sullivan et al. 2006). Additionally, coverage of opioid alternatives such as non-opioid medications, trigger point injections, radio frequency ablutions, and epidural steroid injections can disincentivize the use of opioids in the first place (Schuckit 2016). For members that are already opioid dependent, providing coverage for treatment will reduce the time frame (and associated costs) for individual addiction.

Our last segment of recommendations deals with the actual claims data that Humana tracks in their data warehousing. In the data provided we were able to track if a member had a new provider, but we could not ascertain any other information about the new provider visit, including the provider specialty or the reason for the visit. This information would be useful in understanding why a member has a new provider and in particular for flagging if they are seeing multiple providers for the same medical condition. In addition, when a prescription is rejected we have a general claim reason, but it is vague. The number of prescription rejections a member has in our data is an important predictor of LTOT, but it would be even more informative with the inclusion of additional detail justifying the rejection. Finally, with the rise of remote medicine and mail prescription services, it is important that Humana remains diligent about tracking medical claims as they evolve in this burgeoning space. The healthcare industry is rapidly evolving and Humana must continue to track important member behavior.

We developed recommendations in this section that are derived directly from our model, and intentionally focused on recommendations that are relatively easy to implement in terms of operational and financial efficiency. Whatever recommendations Humana decides to pursue, is important that they develop clear, measurable goals in reducing LTOT incidence so that they can track progress and augment

their strategy if necessary. Based on our work this past month, the unique and arguably most impactful role that insurance companies can play in combating LTOT incidence is in large-scale information tracking, monitoring, and dissemination. Putting the right data in the right hands is the first step to dramatically reducing LTOT for Humana members and beyond.

6 CONCLUSION

In transforming Humana's longitudinal claims data to opioid-naive event level data, we identified that a random forest model could predict LTOT patient outcomes at the time of opioid naive prescription with an AUC of .916. Comparing this model to random chance and a majority class baseline model, it is clear that our model provides meaningful predictions of a Humana members likelihood of an LTOT outcome.
The variable importances of our model offered us interpretable insights that allowed us to formulate recommendations related to sharing information with prescribers, Humana-lead opioid policies, and data tracking practices. We believe that this model's ability to predict outcomes at the patient level and explain more general drivers of LTOT outcomes can assist Humana in combating the prevalence of long term opioid therapies and caring for the health of their members to the best of its ability.

7 APPENDIX

Full Table of Features:

Category	Feature	One Time	All Time	6 Months	1 Year	2 Year
Prescription	# of opioid prescriptions		X	X	X	X
Prescription	Total days of opioid pills prescribed		X	X	X	X
Prescription	Total days of opioids prescribed at opioid naive event	X				
Prescription	Number of different opioids prescribed at opioid naive event	X				
Prescription	# of “OPIOID AGONISTS” prescribed		X			
Prescription	# of “OPIOID COMBINATIONS” prescribed		X			
Prescription	# of “OPIOID PARTIAL AGONISTS” prescribed		X			
Prescription	# of “OPIOID ANTAGONISTS” prescribed		X			
Prescription	# of “ANTIPERISTALTIC AGENTS” prescribed		X			
Prescription	# of opioids prescribed on naive event in “OPIOID AGONISTS” class	X				
Prescription	# of opioids prescribed on naive event in “OPIOID COMBINATIONS” class	X				
Prescription	# of opioids prescribed on naive event in “OPIOID PARTIAL AGONISTS” class	X				
Prescription	# of opioids prescribed on naive event in “OPIOID ANTAGONISTS” class	X				
Prescription	# of opioids prescribed on naive event in “ANTIPERISTALTIC AGENTS” class	X				
Prescription	# of prescriptions for pain		X			
Prescription	# of prescriptions for other		X			
Prescription	# of prescriptions for each opioid type (these are all different variables)		X			
Prescription	Opioid naive prescription was for oral morphine	X				
Prescription	Count of oral morphine prescriptions		X			
Prescription	Opioid naive prescription was for an opioid patch	X				
Prescription	Count of opioid patch prescriptions		X			
Prescription	Opioid prescription MME at Day 0	X				

Prescription	Opioid prescription Max MME		X			
Prescription	Opioid prescription Average MME		X			
Prescription	Count of prescriptions for “psych-anx”		X			
Prescription	Count of prescriptions for “psych-dep”		X			
Prescription	Count of prescriptions for “psych”		X			
Prescription	Count of benzodiazepines prescriptions		X			
Prescription	Total number of non opioid prescriptions		X			
Prescription	Count of NSAID prescriptions		X			
Prescription	Number of rejected claims		X			
Prescription	Number of rejected claims within each statues code (these are all different variables)		X			
Prescription	# of mail prescriptions		X	X	X	X
Diagnosis History	Number of fully paid claims		X			
Diagnosis History	Count of diagnoses containing bipolar		X			
Diagnosis History	Count of diagnoses containing anxiety		X			
Diagnosis History	Count of diagnoses containing psychosis or psychotic		X			
Diagnosis History	Count of diagnoses containing suicide		X			
Diagnosis History	Count of diagnoses containing self-harm		X			
Diagnosis History	Count of diagnoses containing pain		X			
Diagnosis History	Count of diagnoses containing depress		X			
Diagnosis History	Count of diagnoses containing tobacco abuse		X			
Diagnosis History	Count of diagnoses containing alcohol abuse		X			
Diagnosis History	Count of diagnoses containing cocaine abuse		X			
Diagnosis History	Count of diagnoses containing amphetamine abuse		X			

Diagnosis History	Count of diagnoses containing opioid abuse		X			
Diagnosis History	Count of diagnoses containing unspecified abuse		X			
Diagnosis History	Count of diagnoses containing schizophrenia		X			
Diagnosis History	# of CAD diagnoses		X	X	X	X
Diagnosis History	# of diabetes diagnoses		X	X	X	X
Diagnosis History	# of hypertension diagnoses		X	X	X	X
Diagnosis History	# of CPD diagnoses		X	X	X	X
Diagnosis History	# of CHF diagnoses		X	X	X	X
Surgery	# of surgeries in hospital emergency room		X			
Surgery	# of surgeries in other/unknown		X			
Surgery	# of surgeries in home		X			

8 REFERENCES

Are Insurers Doing Their Part to End America's Opioid Epidemic? - Pacific Standard. (n.d.). Retrieved October 15, 2019, from https://psmag.com/social-justice/are-insurers-doing-their-part-to-end-americas-opioid-epidemic

Committee on Homeland Security and Governmental Affairs. (2017). *COMBATTING THE OPIOID EPIDEMIC: A REVIEW OF ANTI-ABUSE EFFORTS IN MEDICARE AND PRIVATE HEALTH INSURANCE SYSTEMS* (pp. 1–30).

Diep, F. (2018, October 7). Are Insurers Doing Their Part to End America's Opioid Epidemic? - Pacific Standard. Retrieved from Pacific Standard Magazine website: https://psmag.com/social-justice/are-insurers-doing-their-part-to-end-americas-opioid-epidemic

How to tell if a loved one is abusing opioids - Mayo Clinic. (n.d.). Retrieved from https://www.mayoclinic.org/diseases-conditions/prescription-drug-abuse/in-depth/how-to-tell-if-a-loved-one-is-abusing-opioids/art-20386038

Kaye, A. D., Jones, M. R., Kaye, A. M., Ripoll, J. G., Jones, D. E., Galan, V., … Manchikanti, L. (n.d.). Prescription Opioid Abuse in Chronic Pain: An Updated Review of Opioid Abuse Predictors and Strategies to Curb Opioid Abuse (Part 2). *Pain Physician*, (2017), 24.

Lautieri, A. (2019, September 25). Benzos and Opiates: Mixing Xanax, Oxycodone, Vicodin & Valium. Retrieved from American Addition Centers website: https://americanaddictioncenters.org/prescription-drugs/dangers-of-mixing

Phillips, J., Ford, M., & Bonnie, R. (2017, July 13). Trends in Opioid Use, Harms, and Treatment - Pain Management and the Opioid Epidemic. Retrieved from National Center for Biotechnology Information website: https://www.ncbi.nlm.nih.gov/books/NBK458661/

Schuckit, M. (2016). Treatment of Opioid-Use Disorders | NEJM. *The New England Journal of Medicine*, *375*. Retrieved from https://www.nejm.org/doi/full/10.1056/NEJMra1604339

Sullivan, M., Edlund, M., & Zhang, L. (2006). Association Between Mental Health Disorders, Problem Drug Use, and Regular Prescription Opioid Use | Psychiatry | JAMA Internal Medicine | JAMA Network. *JAMA Internal Medicine*, *166*(19). Retrieved from https://jamanetwork.com/journals/jamainternalmedicine/article-abstract/411106

Trescot, A., Datta, S., & Lee, M. (2008). Opioid Pharmacology. *Opioid Pharmacology*, (11), 22.