Classifying Opioid Prescription Using Machine Learning Techniques*

CAPP 30254—Machine Learning for Public Policy

Wesley Janson
wrjanson@uchicago.edu

Matt Kaufmann
mkaufmann1@uchicago.edu

Piper Kurtz
kurtzp@uchicago.edu

Angela The
angelathe@uchicago.edu

Eujene Yum
eujeneyum@uchicago.edu

June 2, 2022

Abstract

With nearly one million since 1999, opioid deaths have quickly become one of the leading causes of death, particularly for young people in the United States (CDC NCHS, 2020). Determining who is likely to abuse or misuse opioids can be difficult, but is imperative to fighting this national health crisis. We set out to answer a related question, one that must be first answered —who is more likely to be *prescribed* opioids? In this paper, we aim to predict whether an individual will be prescribed opioids based off observed characteristics using survey data from the 2014-2019 Medical Expenditure Panel Surveys. We fit 5 different machine learning models, all with varying, but promising degrees of success. We find that the random forest algorithm has the highest accuracy of classification, but the decision tree performs the best with respect to recall.

*We would like to thank Jacob Jameson for his helpful comments and shared code.

1 Introduction

Becoming increasingly more prevalent in the recent decades, opioid deaths are now a major cause of death. In 2019, more people died from opioids (49,860) than from motor vehicle accidents (38,800) or breast cancer (42,281) (Centers for Disease Control and Prevention, National Center for Health Statistics 1999–2019). Rooting out the characteristics of individuals who are more likely to be prescribed these potentially-addicting medications could be the first step in addressing, and remedying the ongoing opioid crisis. There are many low-cost and effort interventions that could prove useful in preventing opioid misuse and addiction.

With the proliferation of big data and greater computational power, recent studies of medical survey data have began to implement machine learning algorithms. These studies primarily rely on the Center for Disease Control's National National Survey on Drug Use and Health. We chose to use the Medical Expenditure Panel Survey (MEPS) instead for its comprehensive surveying of families and individuals, their medical providers, and employers across the United States. This thoroughness allows for individual responses to be cross-validated, which can help avoid erroneous responses—for example, if a patient misreports a prescription, their physician may be able to correct it.

2 Data & Relevant Literature

We utilize data from the Medical Expenditure Panel Survey (MEPS), a set of large-scale surveys of families and individuals, their medical providers, and employers across the United States. We specifically use the following three datasets:

1. Prescribed Medicine (PMEDS): This data contains information on what kind of medication a person was prescribed along with whether the prescription was due to an injury. This dataset contains the information on whether a person has been prescribed opioids or not.

2. Medical Conditions (MC): The MC dataset is a condition-level dataset that contains information on the medical conditions of each individual.

3. Household full-year (HC): This data contains information on demographics, health status, access to care, employment, quality of care, health insurance, and other socio-demographic information.

2.1 Data Cleaning & Basic Exploratory Statistics

Using the three individual datasets, we are able to merge information relevant to our modeling. In the prescription medication dataset, we collapse all prescriptions by its individual and create variables for total prescriptions, number of non-opioid prescriptions, and an indicator variable if there were opioids prescribed. On aggregate, $\approx$3.4% of prescriptions were for opioids and $\approx$17% of users with prescriptions were prescribed opioids as many users have multiple prescriptions.

In the medical conditions data, we aggregate the number of medical conditions for each individual. We also add an indicator variable that specifies whether or not any conditions an individual had was due to injury. Combining the 2014 to 2019, data we have 101,680 unique individuals. Approximately 25.3% of individuals in a given year with a medical condition had at least one condition caused by injury, likely a common cause of opioid prescription. Additionally, the number of conditions is right-skewed with an average of 4.57 conditions per person in a year (with a median of 3).

The household full-year dataset will provide many of the features going into our model such as demographics, whether the individual has access to health care, employment, health status, etc. We find that around 53% of respondents are currently employed, with an average household income of $72,025 (median-$53,000). Furthermore, around 95% of respondents have health insurance in some capacity (with the largest share being private insurers).

Our final dataset has 65,871 observations. Though a small sample of those surveyed are

surveyed multiple times in the sample, we treat each instance of a surveyed individual separately, treated the data as a cross-section. We have 32 potential features and a column indicating whether the patient was prescribed opioids, which will act as our variable of interest.[1]

2.2 Relevant Literature

As mentioned before, most studies using machine learning to analyze opioid misuse, abuse, or prescription primarily rely on the Center for Disease Control's National Survey on Drug Use and Health. Han et al. (2020) fit four different models to determine which can best predict adolescent opioid abuse. They find that a distributed random forest model showed the best performance in prediction, closely followed by penalized logistic regression, gradient boosting machine, and artificial neural networks. Their findings suggest that machine learning techniques can be a promising technique especially in the prediction of outcomes with rare cases (i.e., when the binary outcome variable is heavily lopsided) such as adolescent opioid misuse, or in our case, prescription.

Furthermore, Clarke et al. (2014) used a multivariable logistic regression to determine factors that contribute to one or more prescriptions of opioids 90 days after major surgery. The study found that lower income individuals and individuals who used prescription medications before their operation were more at risk for opioid abuse, as well as an inverse correlation between age and probability of opioid abuse. Type of surgery was also a major indicator of likelihood to abuse opioids, and it was found (in other studies) that preoperative fear and anxiety (possible to link to mental health) are correlated with post-surgical pain, which contributes to continued opioid use. Although their study is limited to people ages 65 years or older, their findings suggest important features that should be considered when running our models.

A more comprehensive review by Bharat et al. (2021) of predictive modeling techniques to evaluate overdoses among those with opioid use disorders. They found that deep neural network

[1]The exhaustive list of variables in the final dataset can be seen in Appendix A.5.

and gradient-boosting machine models had the highest measures of discrimination performance, while the logistic regression model had the lowest measures.

3 Models

In total, we fit 5 different models in an attempt to predict opioid prescription in the dataset. Our baseline approaches used logistic regression and a decision tree, the results from both of which were promising, but left room for improvement. Our logistic regression didn't allow for non-linearly separable relationships, and our decision tree's prediction accuracy was low compared to that of logistic regression or our later models. Venturing into more sophisticated machine learning models, we next tried to classify prescriptions using Random Forest, Neural Net, and Ada Boost algorithms.

However, before fitting any models, our first step was to ensure relevant variables did not suffer from multicollinearity by empirically testing. Using variance inflation factor analysis (VIF), a widely accepted off-the-shelf test for multicollinearity among variables, we identify variables that show a strong correlation with others that could contaminate estimation. The idea behind VIF is simple: calculate the ratio of the model variance of estimating some feature variable i in a model that includes multiple other features by the variance of a model constructed using only the i^{th} feature. This provides us with a simple representation of how much the variance of an estimated regression coefficient is increased due to collinearity. A feature variable with a VIF > 5 is considered highly collinear, and its inclusion in a model would contaminate estimation. We make an algorithm that performs VIF analysis on each potential feature variable. From there, if there is at least one variable with a VIF > 5, we eliminate it from our list of potential features. This process is repeated until no variable has a VIF exceeding 5. We found that only two variables suffered from multicollinearity, leaving us with 30 feature variables for our models. Upon completing collinear variable elimination through VIF, we moved on to model construction.

To predict if an individual will receive an opioid prescription or not we developed two separate binary classification models as baselines, a logistic regression model, and a decision tree, using the variables remaining after the VIF analysis. After separating into training, validation, and testing data, we used `sklearn`'s logistic regression package and the `lbfgs` solver to train a logistic regression model. We found our initial linear regression model did not converge on the training set, which could lead to high variance within our model. To try and address this, we chose to scale our data to match a standard normal Gaussian distribution, centered at zero with unit variance ($X \sim \mathcal{N}(0,\, 1)$). Separating the data again and conducting another logistic regression on the normalized data, using the $L2$ regularization parameter, we found that our model unfortunately still did not converge. We then proceeded to train a decision tree model utilizing `sklearn`'s `DescisionTreeClassifier` package.

After successfully running our baseline models, we decided to develop them further to see whether we'd achieve different results with more refined models. For our logistic regression, we moved from a single-layer network to a multi-layer network. We tested the network using `sklearn`'s `NeuralNetworkClassifier` package. We ran the model under various conditions including multiple layer options such as: (100, 50, 25), (10, 5, 2), (25, 25, 25), (10, 9, 8), (10, 5), and (100, 25). For each potential structure, we tested three activation functions (relu, tanh, and logistic). Within each activation function, we tested a variety of learning rates, including 0.0001, 0.001, and 0.01. Additionally, we capped iterations for each model at either 100, 200, or 500 epochs and tested both stochastic gradient-descent and adam, which is a `sklearn` option that is a stochastic gradient-based optimizer proposed by Kingma and Ba (2014). It is unlikely that a single-layer neural network would be able to linearly separate the data on opioid prescription, so our theory was that with a 2-or 3-layer model (two hidden and one output), we should expect to find the global minimum of the loss function and find a linearly separable network. Having a lower number of epochs such as 100 or 200 oftentimes led to the model failing to converge. This model is also necessary as we continue to reconsider variable inclusion in the model; increasing

the dimensions of the attributes necessitates a multi-layer network. We also institute an adaptive learning rate on stochastic gradient descent to balance accuracy and volatility.

To expand upon our decision tree, we developed multiple random forest models utilizing `sklearn`'s `RandomForestClassifier` package. We ran models with different tree counts of 10, 50, 100, 250, and 500. Our hope is that ensemble learning will help us increase the accuracy of our decision tree, although it may decrease interpretability. Although it is difficult to determine the appropriate number of trees to include in the model, we attempt to minimize it (while not compromising accuracy) to limit over-fitting.

Finally, we utilized `sklearn`'s `AdaBoostClassifier` package and tested ran multiple models for learning rates of 0.01, 0.05, 0.1, and 0.2 with 100, 200, 500, and 1000 classifiers. We will explore the results of these models in the next section.

4 Results

When evaluating the accuracy of our models, we first had to define what measure of accuracy to prioritize. We chose to track four different types of accuracy across the variations of our five models: classification accuracy, precision, recall, and F1 score. Classification accuracy is defined as the fraction we correctly labeled a data point. Precision is labeled as the fraction our model, when labelling a data point as positive, was correctly doing so (a measurement to check for the model's ability to correctly avoid false positives). Recall is correct positives over the sum of correct positives and false negatives, meaning it looks at the model's ability to find the true positives. Finally, F1 is the mean score of recall and precision, which allows us to compare the two metrics together.

Given that we are interested in who is most likely to be prescribed opioids, we chose to focus on recall. When considering risk assessment and likelihood to be prescribed opioids, we prefer false positives over false negatives, given this can be used to help with risk assessment

for addiction and it's better to be conservative in this case. Looking at the accuracy results for our validation dataset after training the models, we found a wide range of values in our different accuracy measurements:

Table 1: Validation Set Accuracy Metrics

Model Type	Accuracy	Precision	Recall	F1
Logistic Regression	82.2%	59.1%	10.1%	17.2%
Decision Tree	73.6%	30.0%	32.8%	31.4%
Random Forest	83.0% - 83.5%	65.1% - 75.7%	15.1% - 16.9%	25.0% - 26.8%
Neural Network	77.8% - 82.3%	0.0% - 63.2%	0.0% - 23.7%	0.0% - 28.8%
Ada Boost	82.8% - 83.4%	68.7% - 100.0%	6.8% - 17.4%	12.7% - 27.8%

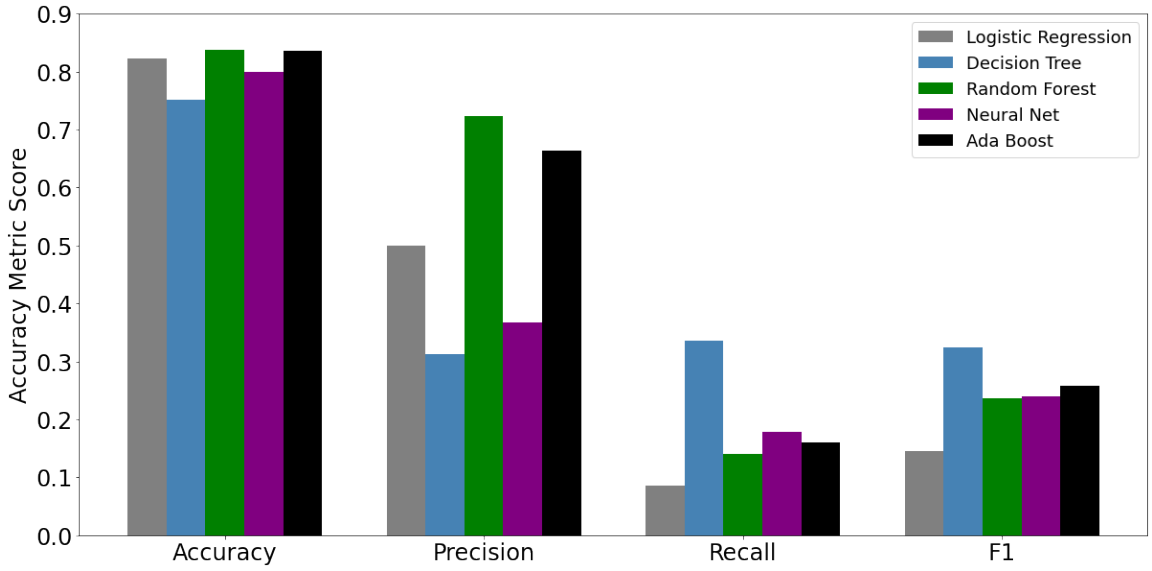
As such, for our final models, we took the individual model that resulted in the highest recall on our validation data to use on our test data. We looked at the highest recall models, and were surprised that Ada Boost performed best with a higher learning rate, but otherwise the specifications for random forest and neural net were generally in line with expectations. The final models utilized were:

1. Logistic regression with 1000 maximum iterations (still did not converge)
2. Basic decision tree classifier
3. Random forest with 1000 estimators
4. Neural network with layers of (100, 50, 25), relu activation function, learning rate of 0.0001, adam solver method, and 500 maximum iterations
5. Ada Boost classifier with 1000 estimators and a learning rate of 0.2

Of note, the random forest model had a wide range of variables in its various stumps, with no variable appearing more than 14% of the time (non-opioid prescription binary variable). See Appendix A.2 for the top ten most utilized variables.

Our final models all perform reasonably well on accuracy, with decision tree the lowest at 75.1% and random forest the highest at 83.8%. The same bookends occurred on the precision metric, with decision tree the lowest at 31.3% and random forest the highest at 72.4%. Interestingly, recall flipped this trend with decision tree the highest at 33.5% but now logistic regression the lowest at 8.5%. Finally, F1 score had decision tree as the best at 32.4% and logistic regression the lowest at 14.6%. A graphical summary of these results can be found below:

Figure 1: Comparing accuracy metrics across final model types.



While random forest was the best at avoiding false negatives (precision), it was near the bottom of identifying true positives (recall). Additionally, given that we care the most about recall accuracy, the decision tree model would be best to utilize going forward, as it was nearly $2\times$ better than the second best model (neural network) in this metric even though it has the lowest overall accuracy and precision metrics.

5 Conclusion

When working with the data, we found that although most of our models reported relatively high accuracy rates, this didn't lead to high values for other measurements of correctness. Determining what measure of accuracy for our project was an important step, and took contextual consideration of the problem, rather than any pure conclusions from the data. Recognizing the importance of recall in this field, we found that among our models, our highest recall value was 33.5% from the decision tree model.

Certainly of note, we saw that a more complex model is not necessarily better than others. When measuring recall, our simple decision tree model performed the best, even when compared to multi-layer neural networks and random forest models. Our more complex models tended to have higher accuracy rates, but oftentimes, this came at the cost of significantly lower recall and F1 accuracy.

Potential next steps for this project would include expanding the dataset to include more observations in order to create more robust models. For example, $\approx$17% of our dataset was prescribed opioids, which is higher than the national average. Having a dataset that was expanded to be more reflective of the population would help when expanding conclusions drawn from the model to the population. Adding additional attributes to the dataset, such as family history of addiction, a community connectedness value, and more in depth mental health data, among other things, would be valuable when predicting opioid prescription. We would also explore using non-greedy decision tree algorithms to ensure an optimal tree as we have seen the highest recall rate with the decision tree model. For further work in the field, we would consider conducting a study to construct a model that predicted opioid abuse, which in concordance with our opioid prescription model, could be used to identify at risk individuals.

Bibliography

Bharat, C., Hickman, M., Barbieri, S., & Degenhardt, L. (2021). Big data and predictive modelling for the opioid crisis: Existing research and future potential. The Lancet Digital Health, 3(6), e397-e407.

Centers for Disease Control and Prevention, National Center for Health Statistics. 1999–2019. "Multiple Cause of Death." CDC Wonder Online Database, released in 2020. Data are from the Multiple Cause of Death Files. http://wonder.cdc.gov/mcd-icd10.html.

Clarke, H., Soneji, N., Ko, D. T., Yun, L., & Wijeysundera, D. N. (2014). Rates and risk factors for prolonged opioid use after major surgery: population based cohort study. BMJ, 348.

Han, D. H., Lee, S., & Seo, D. C. (2020). Using machine learning to predict opioid misuse among U.S. adolescents. Preventive Medicine, 130, 105886.

Kingma, D.P., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

A Appendix

A.1 Supplementary Final Model Accuracy Metrics

Table 2: Test Set Accuracy Metrics

Model Type	Accuracy	Precision	Recall	F1
Logistic Regression	82.2%	50.0%	8.5%	14.6%
Decision Tree	75.1%	31.3%	33.5%	32.4%
Random Forest	83.8%	72.4%	14.1%	23.6%
Neural Network	79.9%	36.7%	17.8%	24.0%
Ada Boost	83.6%	66.3%	16.0%	25.7%

A.2 Graphical Results

Supplementary graphical analysis seen below shows further results from our models.

Figure 2: Most important features of random forest model

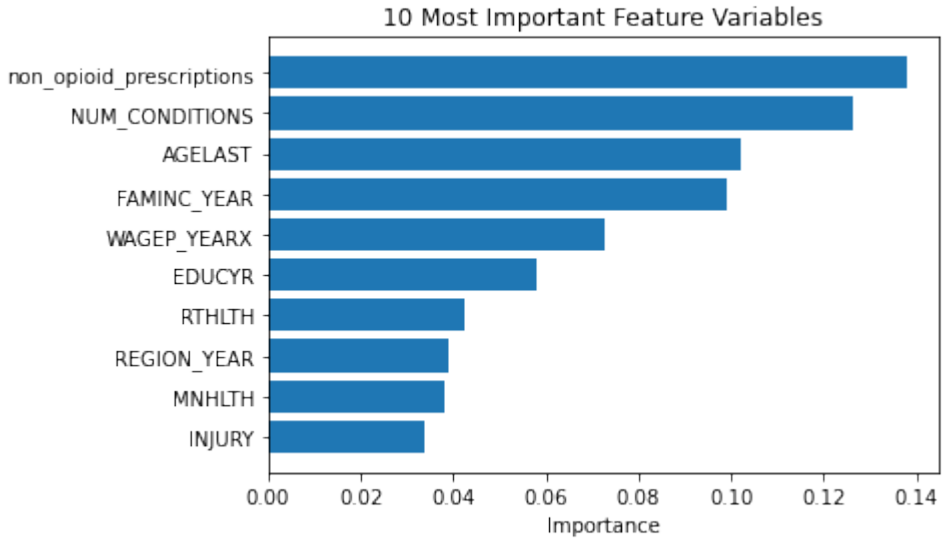
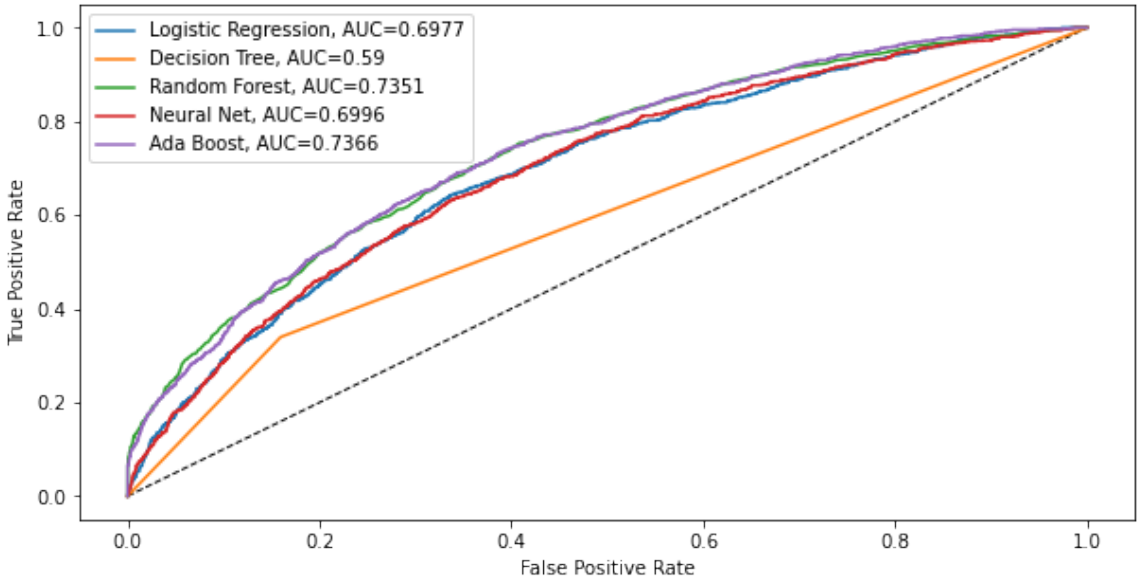


Figure 3: ROC Curve



A.3 Logistic Regression Calculation

Logistic regression works to fit a linear model to data with a binary response variable.

$$P(Y = 0|X') = \frac{1}{1 + exp[\beta_0 + \sum_j \beta_j x]}$$

$$P(Y = 1|X') = \frac{exp[\beta_0 + \sum_j \beta_j x]}{1 + exp[\beta_0 + \sum_j \beta_j x]}$$

Where Y is the outcome variable of whether one was prescribed an opioid and X' is a vector of j feature variables. We set out to minimize the negative log likelihood function with $L2$ regularization

$$\text{NLL}_{L_2}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \left[Y_i \log \sigma(\boldsymbol{\beta}^T \mathbf{x}^{(i)}) + (1 - Y_i) \log(1 - \sigma(\boldsymbol{\beta}^T \mathbf{x}^{(i)})) \right] + \lambda \sum_{k=1}^{j} \beta_k^2$$

A.4 Decision Tree Splitting

Using the decision tree model, we split to maximize information gain (IG) at every node. This uses the concept of $entropy$, denoted as $H(X)$, which can be described as measuring the randomness of the information being processed at a node, where p_c is the fraction of examples in class c.

$$H(X) = -\sum_{c} p_c \; log_2(p_c)$$

$$IG(X_{p,i}) = H(X_p) - \frac{|X_{i,left}|}{|X_p|} H(X_{left}) - \frac{|X_{i,right}|}{|X_p|} H(X_{right})$$

A.5 Relevant Variables

Table 3: Variables in Final Data

Variable	Description
$OPIOID\_PRESCRIBED\_AT\_ALL^*$	Binary variable whether patient was prescribed at least one opioid.
$YEAR$	Year survey was taken.
$REGION\_YEAR$	Census Region of residency at the end of year survey is taken.
$AGELAST$	Person's age last time eligible for survey.
SEX	Sex.
$RACETHX$	Race/Ethnicity.
$MARRY\_YEARX$	Marital status at end of calendar year survey was taken.
$EDUCYR$	Years of education attained when first entered MEPS.
$BORNUSA$	Binary variable whether person was born in the US.
$FOODST\_YEAR$	Binary variable of whether food stamps received in the past year.
$TTLP\_YEARX$	Person's total income in survey year.
$FAMINC\_YEAR$	Family's total income in survey year.
$POVCAT\_YEAR$	Family income as percent of poverty line - categorical.
$POVLEV\_YEAR$	Family income as percent of poverty line - continuous.
$WAGEP\_YEARX$	Person's wage income in survey year.
$DIVDP\_YEARX$	Person's dividend income in survey year.
$SALEP\_YEARX$	Person's sales income in survey year.
$PENSP\_YEARX$	Person's pension income in survey year.
$PUBP\_YEARX$	Person's public assistance income in survey year.
$ADHDADDX$	Binary variable whether person is diagnosed with ADHD.
$UNINSURED\_ONLY$	Binary variable whether person is uninsured.
$PRIVATE\_ONLY$	Binary variable whether person is covered by private insurance only.
$MEDICAID\_ONLY$	Binary variable whether person is covered by Medicaid only in.
$MEDICARE\_ANY$	Binary variable whether person is covered in any capacity by Medicare.
$MEDICARE\_ADV$	Binary variable whether person is covered by Medicare Advance.
$MEDICARE\_MEDICAID$	Binary variable whether person is covered by Medicare and Medicaid.
$MEDICARE\_PRIVATE$	Binary variable whether person is covered by Medicare and private insurance.
$ACTDTY$	Binary variable whether military full-time active duty.
$RTHLTH$	Perceived health status.
$MNHLTH$	Perceived mental health status.
$EMPST$	Employment status.
$NON\_OPIOID\_PRESCRIPTIONS$	Number of non-opioid prescriptions.
$NUM\_CONDITIONS$	Number of medical conditions.
$INJURY$	Binary variable whether any condition is a result from an injury sustained.

$*$ Variable of interest in models.